Speech Communication 149 (2023) 74–83

Contents lists available at ScienceDirect

Speech Communication

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

Using a single-channel reference with the MBSTOI binaural intelligibility metric

Pierre Guiraud *, Alastair H. Moore, Rebecca R. Vos, Patrick A. Naylor, Mike Brookes

Imperial College London, Department of Electrical and Electronic Engineering, London, SW7 2AZ, United Kingdom

ARTICLE INFO

Keywords:
MBSTOI
Deep learning
Speech intelligibility metric

ABSTRACT

In order to assess the intelligibility of a target signal in a noisy environment, intrusive speech intelligibility metrics are typically used. They require a clean reference signal to be available which can be difficult to obtain especially for binaural metrics like the modified binaural short time objective intelligibility metric (MBSTOI). We here present a hybrid version of MBSTOI that incorporates a deep learning stage that allows the metric to be computed with only a single-channel clean reference signal. The models presented are trained on simulated data containing target speech, localised noise, diffuse noise, and reverberation. The hybrid output metrics are then compared directly to MBSTOI to assess performances. Results show the performance of our single channel reference vs MBSTOI. The outcome of this work offers a fast and flexible way to generate audio data for machine learning (ML) and highlights the potential for low level implementation of ML into existing tools.

1. Introduction

The intelligibility of a noisy speech signal can be measured directly using listening tests. However, such tests may be lengthy and difficult so that it is often preferred to estimate intelligibility using one of the many available instrumental or computed intelligibility metrics. The most reliable current metrics are intrusive, meaning that they require the original, clean signal as a Ref. French and Steinberg (1947), ANSI (1997), Hu and Loizou (2008), Kates and Arehart (2014, 2021), Jørgensen and Dau (2011) and Jørgensen et al. (2013).

One of the most popular single-channel intrusive metrics remains the short time objective intelligibility metric (STOI) (Taal et al., 2011). This was later improved, resulting in ESTOI, to handle fluctuating interferers better (Jensen and Taal, 2016), before being extended into a binaural metric by adding an equalisation cancellation (EC) stage in DBSTOI (Andersen et al., 2016). DBSTOI however was found to overestimate intelligibility with spatially distributed interferers when the signal to noise ratio (SNR) was low. This inaccuracy was resolved with the introduction of a modified version, called MBSTOI, in Andersen et al. (2018a) which improves the robustness of the metric by using a modified optimality criterion for determining the EC parameters. The EC stage (Durlach, 1963) has also been used in other binaural metrics (stBSIM in Beutelmann and Brand (2006), Beutelmann et al. (2010) or B-sEPSM in Chabot-Leclerc et al. (2016)) and uses binaural cues to find parameters to align and cancel undesired localised interference.

The recent widespread availability of machine learning (ML) and deep learning (DL) tools (Abadi et al., 2016; Paszke et al., 2019) has led to the creation of many new metrics. Where end-to-end metrics have been created (Pedersen et al., 2020; Cauchi et al., 2019), these are mostly aimed at improving presented techniques (ANIQUE+, STOI-NET, NI-SIP) (Kim and Tarraf, 2007; Zezario et al., 2020; Andersen et al., 2018b). The majority of existing methods implement single-channel metrics although recently some metrics have applied ML to binaural signals such as the binaural speech intelligibility metric (BAPSI) (Rosbach et al., 2021).

While binaural metrics perform well, the need for a binaural clean reference limits their scope of application. Following a sensitivity analysis of EC estimation errors in MBSTOI, it has recently been shown that it is possible to replace this EC estimation in MBSTOI by a DL network and so create a hybrid metric (Guiraud et al., 2022). ML is here integrated at a low level so that most steps of the metric computation remain unchanged. In the present paper, a DL model is used to reproduce the performance of traditional MBSTOI using only a single-channel reference signal rather than a binaural reference. Its performance is assessed in a range of situations using both simulated datasets and an independent open-source dataset of real recordings.

The principles of MBSTOI are presented in Section 2.1 before introducing two new hybrid ML metrics in Section 2.2. Details of the created datasets, DL models and training process are provided in Sections 3, 4.1 and 4.2 respectively. Analysis of the performance on the created datasets is presented in Sections 5.1 and 5.2. Section 5.3 shows the performance of the metrics with the independently recorded dataset. Finally, conclusions and perspectives are given in Section 6.

* Corresponding author.
E-mail address: pguiraud@imperial.ac.uk (P. Guiraud).

https://doi.org/10.1016/j.specom.2023.03.005
Received 30 September 2022; Received in revised form 3 February 2023; Accepted 6 March 2023
Available online 11 March 2023

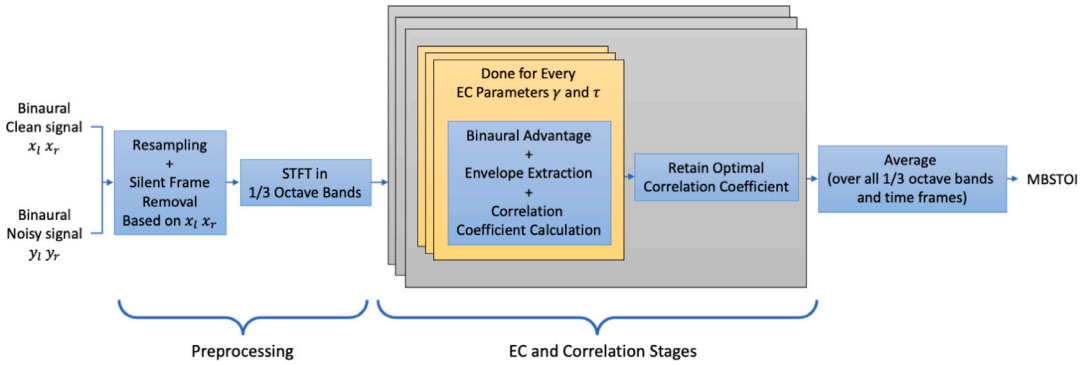

Fig. 1. MBSTOI computation block diagram. The EC stage is performed independently at every time frame and over all third-octave bands. The estimation of the EC parameters consists of a grid-like search and an optimisation criterion determines the retained value for each time and frequency.

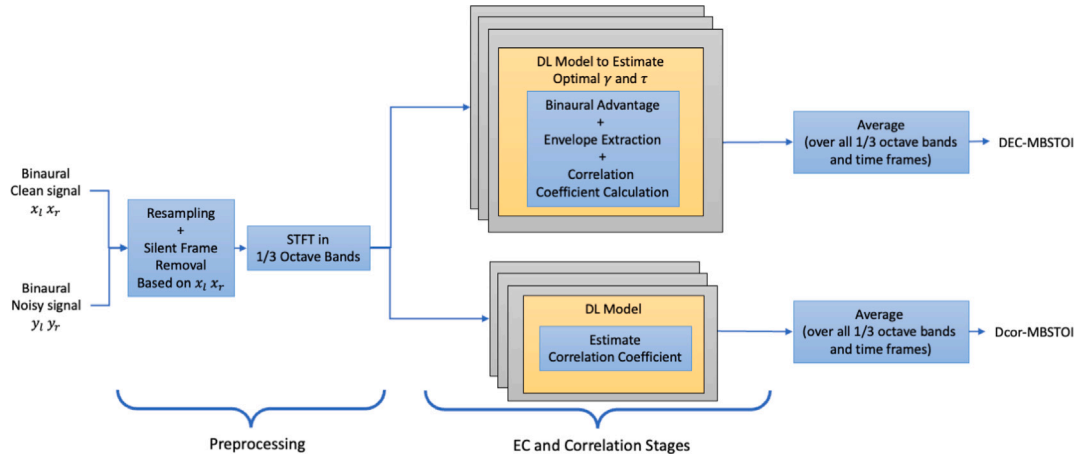

Fig. 2. DEC and Dcor MBSTOI computation block diagram. The preprocessing stage is unchanged from traditional MBSTOI. DEC-MBSTOI uses DL to estimate directly the optimal EC parameters to remove the grid search. Dcor-MBSTOI estimates directly the correlation coefficient of the current time frame and frequency band.

2. MBSTOI, DEC-MBSTOI and Dcor-MBSTOI

2.1. MBSTOI: Principle, EC stage and correlation

The calculation of the MBSTOI binaural intelligibility metric is illustrated in Fig. 1. The inputs to the metric are a noisy binaural signal and a clean binaural signal containing only the target speech components. The signals are resampled to 10 kHz and divided into overlapping analysis frames. After removing frames in which the target speaker is silent, the signals are transformed into the frequency domain using the short-time Fourier Transform (STFT), and the frequency bins aggregated into third-octave bands. The DFT coefficients of both ears are combined to model binaural advantage in intelligibility when there is spatial separation between target and interferer. This EC stage uses parameters which represent the interaural level and time differences (ILD and ITD) of the dominant interferer and are denoted γ and τ respectively. However, since the direction of the source and interferer are not known, an exhaustive grid search of the EC parameters, with resolutions of 1 dB and 20 µs in γ and τ respectively, is performed to determine the potential advantage of having two ears. A single pair of values is selected for each frequency band based on the optimisation criterion and then used to incorporate the binaural advantage into MBSTOI (Andersen et al., 2018a). The correlation between clean and degraded envelopes is then computed and averaged across all time frames and all frequency bands to predict a metric value between 0 and 1.

2.2. DEC-MBSTOI and Dcor-MBSTOI

Within MBSTOI, the clean binaural reference signal is then used for three purposes:

- silent frame removal,
- estimation of the EC parameters, γ and τ,
- calculation of the correlation coefficients.

If the binaural reference were to be replaced by a single-channel reference, this would not significantly influence the silent frame removal stage but the EC parameters and correlation coefficients estimation would need to be modified. In this work, two approaches to using ML in place of the EC stage are investigated. In both cases, the end goal is to be able to replace the binaural reference signal by a single-channel reference signal in the computation of MBSTOI.

Fig. 2 summarises the processing stages of the two metrics. The first metric, called deep equalisation cancellation MBSTOI (DEC-MBSTOI), eliminates the exhaustive search for the optimal values of γ and τ. Instead, a deep neural network (DNN) is trained to estimate the optimal EC parameters and the correlation coefficient is then calculated using the single-channel reference as both the left and right ear channel. The second metric is called deep correlation coefficients MBSTOI (Dcor-MBSTOI). Here, the DNN directly estimates the correlation coefficient and bypasses the use of the EC parameters. The motivation for this metric arises from the difficulty in applying the EC and correlation stages when only a single-channel reference is available.

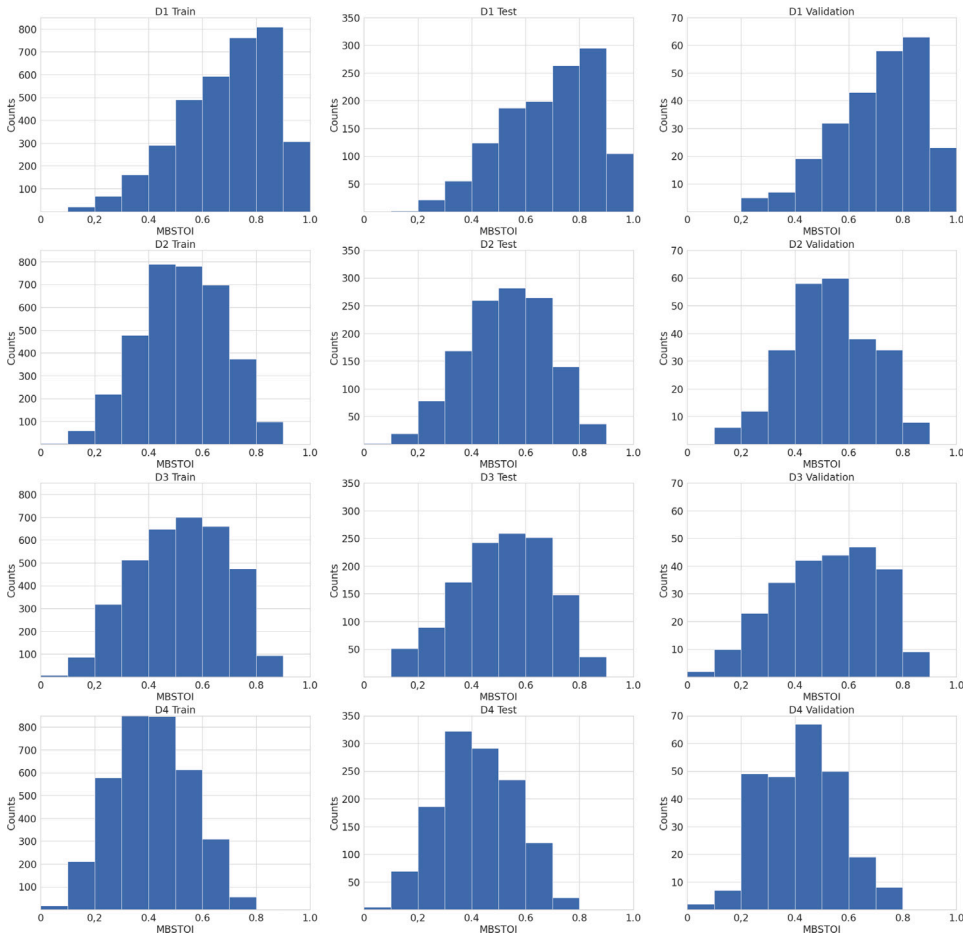

Fig. 3. Histogram distribution of MBSTOI in all four datasets in each Train, Test and Validation split.

Table 1
Dataset details.

	Target	Interferer
Distance (m)	1	1
Level (dB)	60	[50, 51, …, 70]
Azimuth (degrees)	[−30, 0, 30]	[0, 22.5, …, 180]
Elevation (degrees)	[−45, 0, 45]	[−45, 0, 45]

3. Dataset creation

Four datasets with increasing scene complexity were created using `tascar` (Grimm et al., 2019). While these datasets share similar scene features, they were generated independently. Hence, scenes are not equivalent between datasets. The interested reader is invited to look at the `tascar` manual for more details about the various plugins used. For each dataset, 5000 sentences were generated and split randomly into Train, Test and Validation sets comprising respectively 70%, 25% and 5% of the total dataset. The MBSTOI (Andersen et al., 2018b) distribution for each set is shown in Fig. 3. These are the values that the presented hybrid-MBSTOI metrics aim to reproduce.

3.1. Dataset D1: Interfering noise

The first dataset consists of an anechoic scene where a target speaker is located in front of the listener. The speech is randomly chosen from the IEEE speech corpus (Rothauser et al., 1969), UK recordings. This corpus consists of 72 individual male and 72 individual female recordings of up to 10 sentences each. Each token in this corpus is a sentence of less than 6 s duration. An interfering noise signal taken randomly either from the same IEEE speech corpus or from the PNL100 non-speech noise corpus (Hu and Wang, 2010) is played from a position around the listener. If the interfering noise comes from the IEEE corpus, a random different talker (male or female) and sentence is used for the target signal. Sound is played in `tascar` using the `sndfile` plugin.

A simulated listener which incorporates the main features of a measured head related transfer function (HRTF) is located at the origin of the spatial coordinate system. This is implemented using the `receiver hrtf` plugin with default values. The location and level of the target and interferer signal relative to the listener are randomly chosen from the values listed in Table 1. In addition to the binaural HRTF receiver, an omnidirectional microphone (`receiver omni` plugin) is placed at the origin to provide the single-channel reference signal.

3.2. Dataset D2: Diffuse noise

The second dataset adds diffuse noise in addition to the localised interferer from dataset D1. The room considered is still anechoic. The target and interferer signal are chosen in a similar fashion as in D1. The diffuse noise is implemented using `tascar diffuse` plugin which generates a diffuse sound field within a space. Recorded babble noise from a crowded bar is used. Its level is selected from the same range as the interferer in Table 1 but is independently chosen.

3.3. Dataset D3: Reverberant room

The third dataset places the target and interfering noise from dataset D1 in a reverberant environment. The dimensions of the room are varied randomly from a minimum of $2.5 \times 2.5 \times 2.2$ m to a maximum

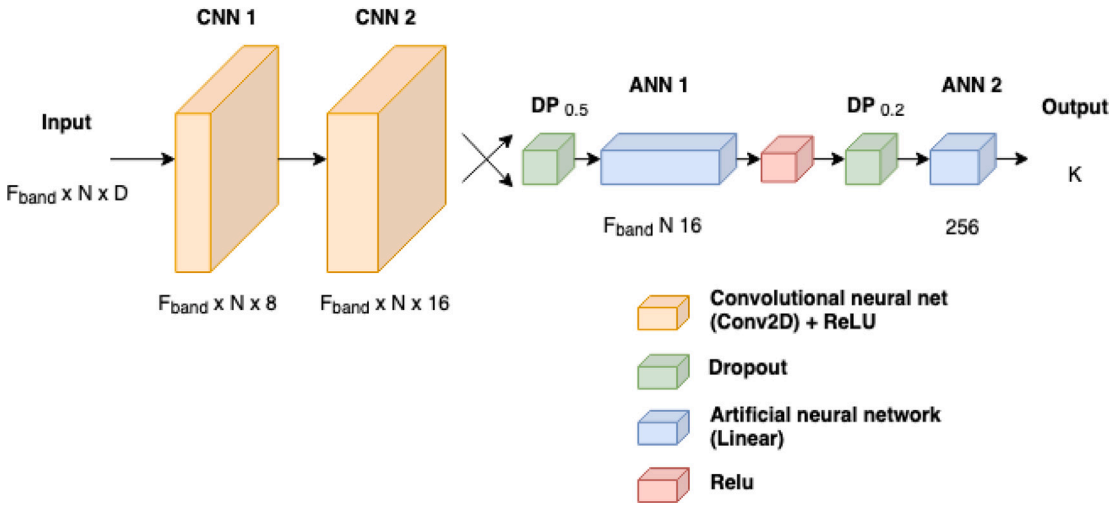

Fig. 4. Schematic of the DNN model used. The legend states in parentheses the exact name of the corresponding layers used in Pytorch (Paszke et al., 2019). $F_{band} \in [2, 45]$ is the number of STFT frequency bins in the current third-octave band, $N = 30$ is the number of time frames, $D = 6$ for a single-channel reference or 8 for a binaural reference, and $K = 2$ for DEC models and 1 for Dcor models.

Table 2
Reverberation time (RT_{60}) calculation for the smallest and largest rooms of the simulations using the Sabine formula.

Room size / absorption coefficient	0.1	0.9
Smallest	0.64 s	0.07 s
Largest	1.79 s	0.20 s

Table 3
ML hyperparameters.

Hyperparameters	Values
Learning rate	0.001
Batch size	16
Max epoch	6
Optimiser	Adam Kingma and Ba (2015)
Criterion	MSE
Early stopping	Patience of 2 on test loss Zhang and Yu (2005) and Yao et al. (2007)

of 10 × 10 × 4 m. The listener is always at the centre of the room. The absorption coefficient of the walls is randomly chosen between 0.1 and 0.9 with a 0.1 step. The early reflections are implemented using `tascar facegroup` plugin and diffuse reverberation is done with a "simple feedback delay network" from the `reverb` plugin. The reverberation time (RT_{60}) is calculated using the Sabine formula $RT_{60} = 0.161V/(S\alpha)$ (Schomer and Swenson, 2002) with V volume, S surface and α average absorption coefficient. The resulting maximum RT_{60} for the smallest and largest room are seen in Table 2. As with dataset D1, there is no diffuse noise.

3.4. Dataset D4: General case

The fourth dataset combines all the `tascar` plugins of the previous datasets. The scenes consist of a target signal with interfering localised noise, diffuse noise and room reverberation. This dataset is intended to represent a diverse and more general case in which to test the metrics. It is stressed that datasets are generated independently, and that sentences do not correspond directly between datasets.

4. Machine learning models and training

4.1. Deep learning models

Fig. 4 shows the architecture of the deep neural networks (DNN) used in both DEC-MBSTOI and Dcor-MBSTOI. Each DNN comprises a combination of convolution neural network (CNN) and artificial neural network (ANN) layers. Details of the hyperparameters of the models are given in Table 3. The input parameters used are the same as for traditional MBSTOI, meaning the spectrograms for each channel of the binaural noisy signal and the binaural or single-channel reference signal. For each single input signal, the size of a spectrogram is the number of STFT frequency bins in the current third-octave band (F_{band}) multiplied by the number of time frames, N, in the analysis window. A different model is trained for each third-octave-band due to the varying number of STFT bins in each band. Each analysis window forms an individual training input.

The use of a binaural or single-channel reference changes the number of input parameters D. Whereas four spectrograms are used for traditional MBSTOI, only three are used in the case of a single-channel reference. In order to preserve the phase information, each input spectrogram is separated into its real and imaginary parts. Hence the number of parameters D of the model is a total of either 8 or 6 real-valued spectrogram components according to whether a binaural or single-channel reference is used.

In Guiraud et al. (2022), more complex structures were tested with no significant improvements in results, these included DNNs with an increased number of layers, nodes and with alternative input parameter representations. The DNN model presented in Fig. 4 is used throughout this paper as it gave accurate MBSTOI estimation and similar computational time to MBSTOI.

All the DNNs were implemented using the PyTorch library (Paszke et al., 2019) and used the Python implementation of MBSTOI provided by the Clarity challenge (Graetzer et al., 2021).

4.2. Training

Fig. 5 details the simulation process and how it is used to train the metrics. Using the available original audio files, `tascar` (Grimm et al., 2019) generates two outputs for all scenes: binaural and single-channel audio recordings at the listener's position. The noisy binaural signals for D1 to D4 are then computed as well as their associated clean

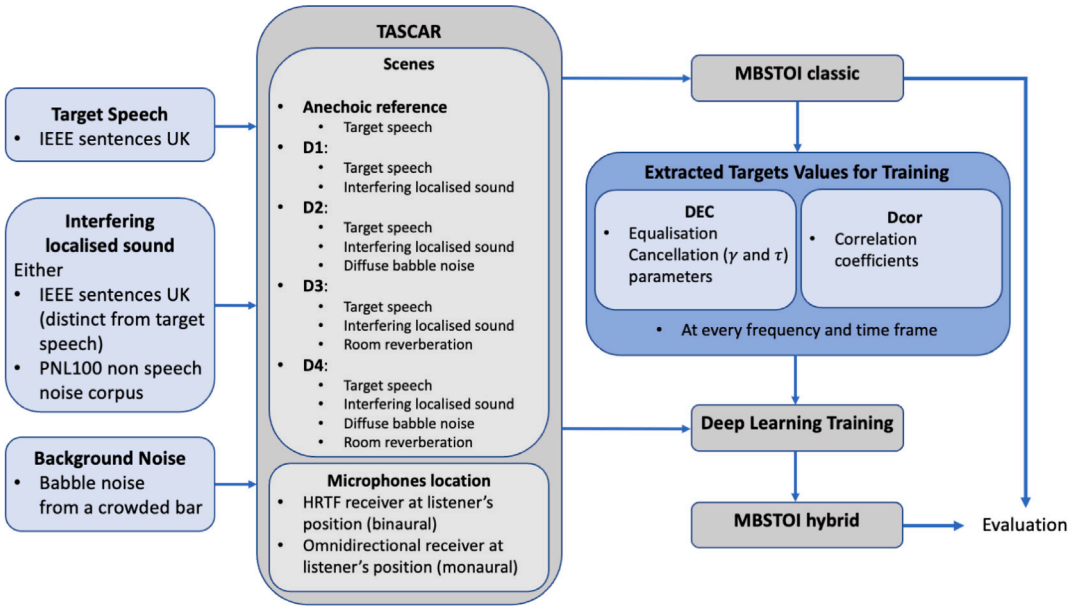

Fig. 5. Block diagram of the construction of the hybrid MBSTOI metrics. `tascar` is used to generate the datasets for training. Traditional MBSTOI is then computed to provide the target outputs being either the EC parameters or the correlation coefficient. D1 to D4 identify alternative datasets with increasing complexity. DEC and Dcor identify alternative portions of the MBSTOI algorithm that are replaced by a neural network.

anechoic reference. Using matching noisy and clean signals, MBSTOI is calculated and the EC parameters and correlation coefficients are extracted. They are then used as target values for DL training using the same signals.

To evaluate the trained metrics, DEC and Dcor MBSTOI results are compared with the target MBSTOI value. It is noted that the new metrics will learn any shortcomings of MBSTOI, since it is used as the training target. In particular, performance results of MBSTOI and the new metrics might differ when compared on human intelligibility data. However, any hypothetical better performance of the hybrid models would be purely coincidental as the DL section is implemented at low level of MBSTOI calculation.

Both metrics are first tested on MBSTOI reproduction using the binaural reference before being trained using the single-channel reference only.

5. Evaluation results

In this section, all presented results were obtained using the Validation sets. In other words, observed data are similar to the Train and Test sets but were never seen before by the ML models during training.

Throughout this section, in order to establish whether two models display statistically significant differences, paired t-tests have been performed on the hybrid-MBSTOI results. However, in order to not break the independent observation hypotheses of paired t-test, the results have been randomly split in two. This process is repeated five times and the average p-value is calculated (Dietterich, 1998; Bouckaert and Frank, 2004; Vanwinckelen and Blockeel, 2012). 5 × 2-fold paired t-test is then applied to every combination of models to all datasets of interest and the results are displayed in Tables 4 and 5. Values in bold are above the 5 % threshold and are considered not statistically different.

5.1. DNN estimation performance

Plots investigating a single representative sentence from D4 can be found in Figs. 6 to 9. They display the accuracy of DEC in estimating γ and τ or of Dcor in estimating the correlation coefficients across frequency bands. The darker, blue, dots correspond to binaural DEC and Dcor while the lighter, orange, dots correspond to the single-channel version.

It is observed in Figs. 6 and 7 for EC parameters estimation, that ML models tend to often output a value close to 0 making for poor estimation. This is confirmed when calculating the Pearson correlation coefficients for all frequency bands. The resulting mean Pearson coefficient is below 0.2 for both γ and τ. However, Fig. 8 shows that the resulting calculated correlation coefficient correlates well with the target ones, with Pearson coefficients of 0.918 for binaural and 0.825 for single-channel. It will be seen in Section 5.2 than this leads to good estimation of DEC-MBSTOI despite poor EC parameter estimation. From a ML point of view, it is likely that many local minima can be found for the EC parameter optimisation that all lead to similar prediction accuracy. This relates well with the unequal importance of

Table 4
Averaged p-values of a 5 × 2-fold paired t-test on the validation data of the various models used across datasets. In bold are the values above 0.05 being not statistically significant.

Compared pair	DEC-b	Dcor-b	DEC-m	Dcor-m
D1/D2	1.28e−02	**6.45e−01**	**4.84e−01**	**3.25e−01**
D1/D3	3.04e−07	1.63e−15	5.08e−08	2.22e−06
D1/D4	1.84e−10	4.10e−12	1.02e−06	3.19e−05
D2/D3	1.79e−03	7.71e−16	3.93e−08	3.19e−08
D2/D4	1.38e−06	3.05e−13	2.91e−07	2.60e−06
D3/D4	1.91e−02	**4.08e−01**	**3.88e−01**	**2.31e−01**
D1/EC = 0	4.08e−08	**2.08e−01**	1.68e−03	2.78e−03
D2/EC = 0	8.47e−11	**3.38e−01**	1.45e−04	1.40e−05
D3/EC = 0	1.37e−15	2.87e−09	**9.06e−02**	**1.58e−01**
D4/EC = 0	4.52e−19	1.48e−07	**5.41e−01**	**5.30e−01**

Table 5
Averaged p-values of a 5 × 2-fold paired t-test on the validation data of the various datasets used across models. In bold are the values above 0.05 being not statistically significant.

Compared pair	D1	D2	D3	D4	Clarity
DEC-b/Dcor-b	1.61e−14	9.62e−22	1.93e−03	1.07e−08	**3.65e−01**
DEC-m/Dcor-m	**7.13e−01**	**3.54e−01**	**5.74e−01**	**4.61e−01**	**6.31e−01**
DEC-b/DEC-m	8.63e−23	7.51e−23	1.03e−13	3.11e−19	6.41e−65
Dcor-b/Dcor-m	1.06e−02	2.41e−04	3.43e−05	7.92e−06	1.67e−57

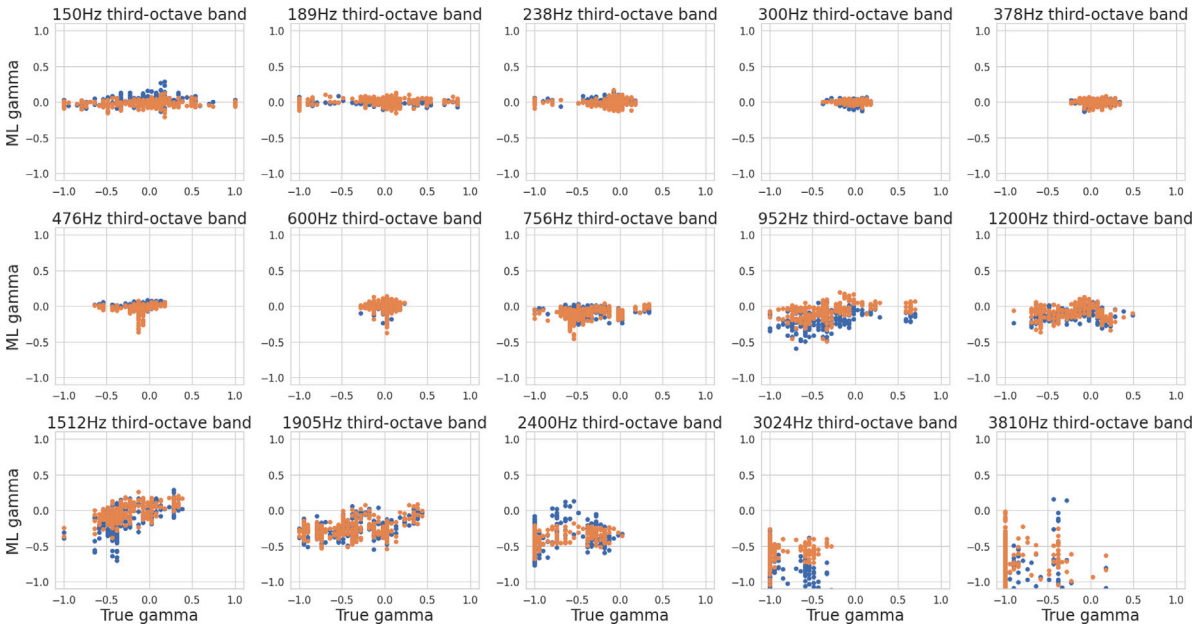

Fig. 6. DL model estimation of gamma across third-octave bands compared to their target value. Title provides the centre frequency. Blue dots correspond to binaural DEC while orange dots correspond to the single-channel version. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

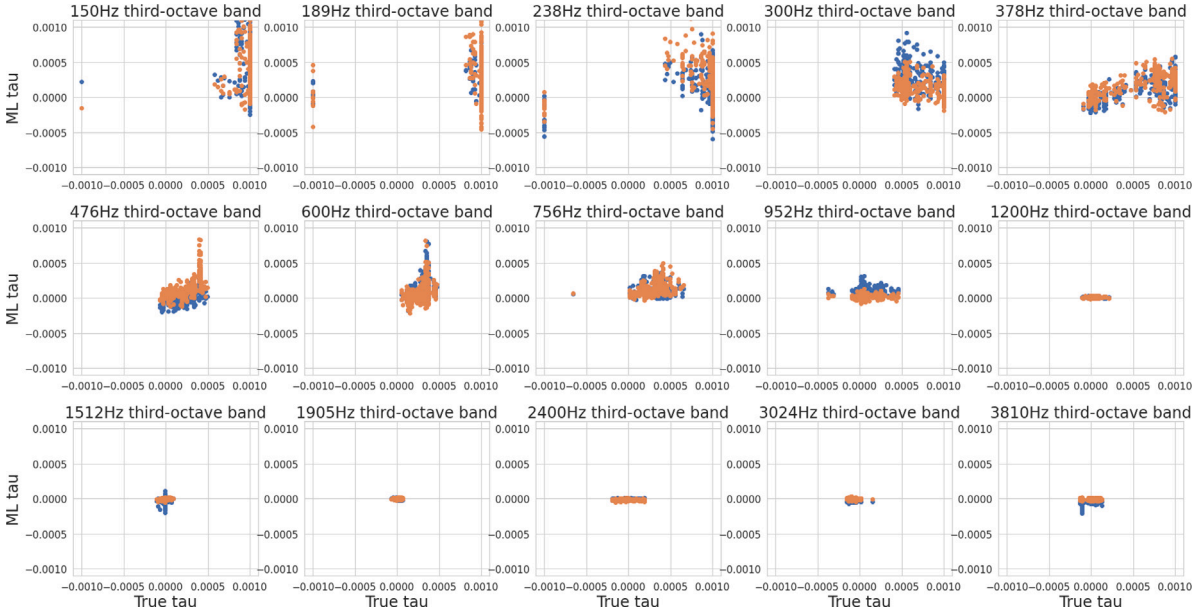

Fig. 7. DL model estimation of tau across third-octave bands compared to their target value. Title provides the centre frequency. Blue dots correspond to binaural DEC while orange dots correspond to the single-channel version. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

EC parameters across frequency bands. More details about the relative importance of accurate estimation of the EC parameters in the MBSTOI calculation are found in Guiraud et al. (2022).

It is seen in Fig. 9 that Dcor models estimate parameters well across frequency bands with similar binaural and single-channel predictions at high frequency. An average correlation of 0.896 and 0.820 is observed for binaural and single-channel reference respectively.

5.2. DEC- and Dcor-MBSTOI

The parameters estimated by the DNNs are subsequently used in the computation of DEC- and Dcor-MBSTOI. The ΔMBSTOI measure is computed as the difference between DEC- or Dcor-MBSTOI with the true MBSTOI.

5.2.1. DEC/Dcor-MBSTOI performance

Fig. 10(a) compares DEC and Dcor when using the anechoic binaural reference, as done in MBSTOI. First it is observed that MBSTOI is almost always underestimated. DEC models estimate MBSTOI with mean absolute deviation below 0.04 across all datasets. It was shown in Guiraud et al. (2022) that this magnitude of error corresponds to a misestimation of the EC parameters of less than 1 dB and 40 µs respectively. Dcor estimation error on the other hand reaches a mean absolute deviation of up to 0.08 for D1 and D2 and drops below 0.04 in D3 and D4. This gain in performance is also observed to a lesser degree in DEC. The addition of reverberation in D3 and D4 seems to help ML models differentiate the target speech from the interferer. This is corroborated in Table 4 where it is shown that statistical difference cannot be established between D1 and D2, or D3 and D4, for Dcor.

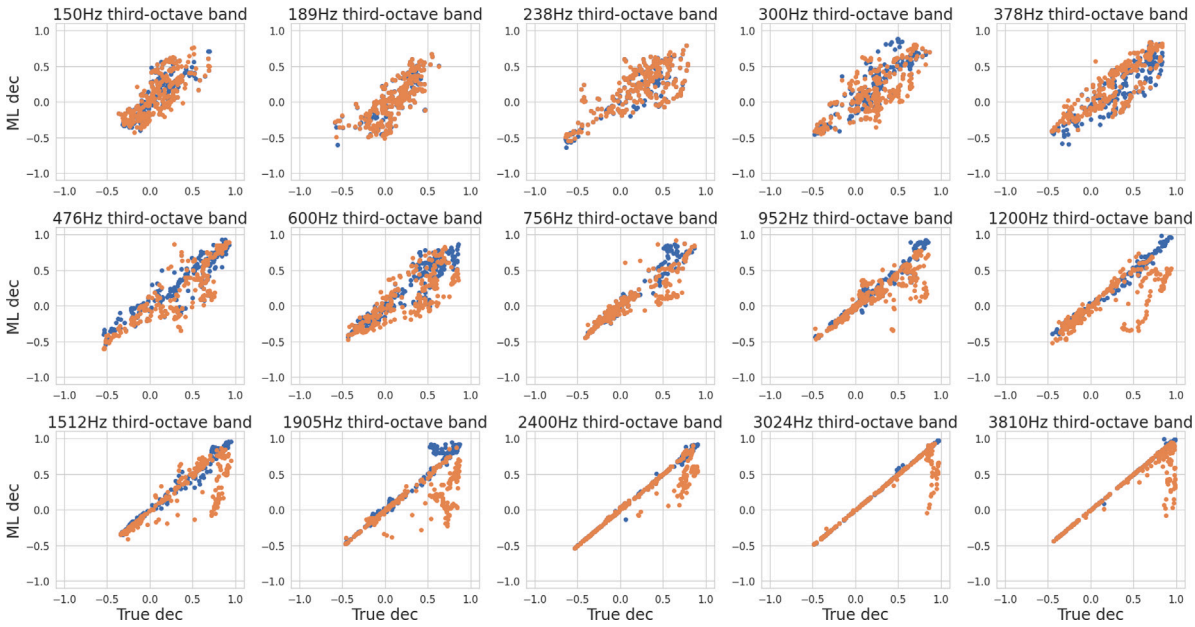

Fig. 8. Calculated correlation coefficient using DL estimated EC parameter compared to their target value across third-octave bands. Title provides the centre frequency. Blue dots correspond to binaural DEC while orange dots correspond to the single-channel version. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

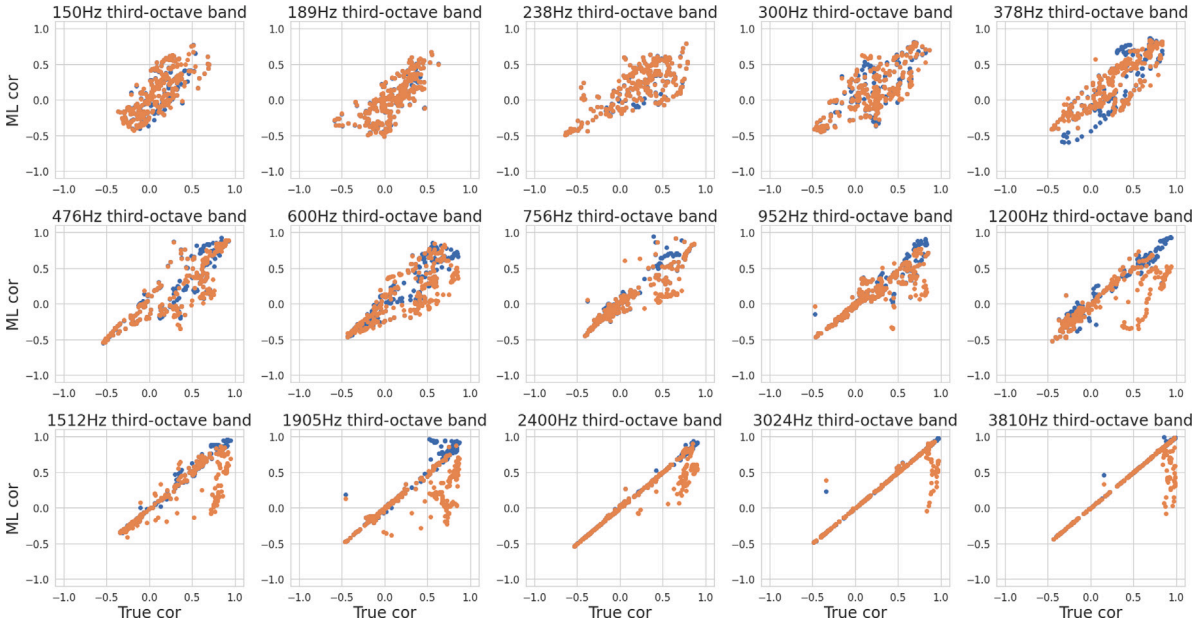

Fig. 9. DL model estimation of correlation coefficient across third-octave bands compared to their target value. Title provides the centre frequency. Blue dots correspond to binaural Dcor while orange dots correspond to the single-channel version. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Only DEC binaural increases in accuracy in a significant way with each datasets. It is also observed that DEC models outperform Dcor models across all four datasets when using a binaural reference, as confirmed by Table 5.

Fig. 10(b) compares results of DEC and Dcor models trained using a single-channel reference signal. Similarly to in Fig. 10(a), MBSTOI is almost always underestimated and performance increases in D3 and D4, which include reverberation, compared to D1 and D2. However, a general decrease in performance is observed compared to binaural. This decrease is more pronounced for DEC leading to comparable performance between DEC and Dcor across the four datasets. Similarly as in Fig. 10(a), Table 4 indicates that statistical difference cannot be established between D1 and D2, or D3 and D4, for both DEC and Dcor. In the most complex situation with D4, the mean absolute deviation of Dcor-MBSTOI is of 0.07 with standard deviation of 0.058.

As an added reference in Fig. 10(a), $\text{MBSTOI}_{EC=0}$ displays MBSTOI variations when γ and τ are manually set to 0. It has a mean absolute deviation of 0.074 and standard deviation of 0.063. According to Table 4 and Fig. 10(a), binaural DEC-MBSTOI models are statistically different and outperform $\text{MBSTOI}_{EC=0}$ across all datasets, while it is only true in D3 and D4 for binaural Dcor-MBSTOI. For models with single channel reference, it is interesting to see that, without reverb (D1 and D2), models are statistically different from $\text{MBSTOI}_{EC=0}$ but perform worse as seen in Fig. 10(b). The increase in performance observed with

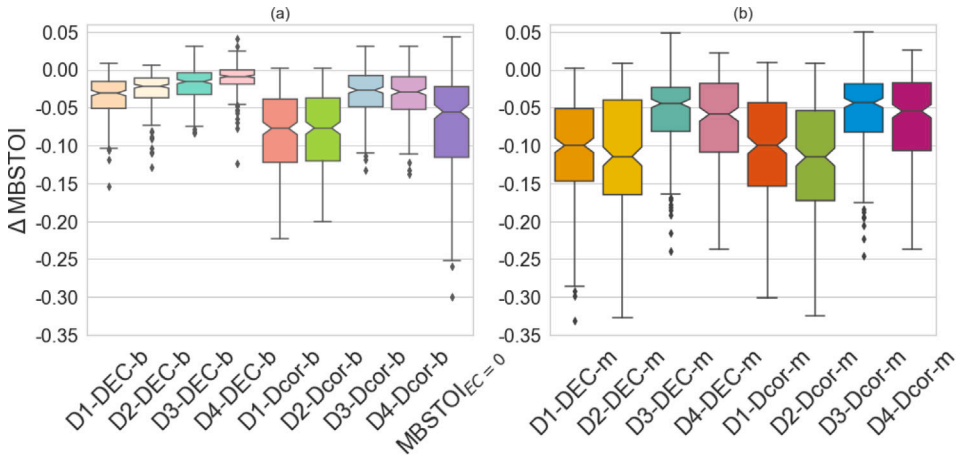

Fig. 10. Box plots representing the difference between the estimated ML MBSTOI and the true MBSTOI value. This comparison is done for all 4 datasets, with DEC and Dcor models, using either a binaural, $-b$, or a single-channel, $-m$, reference. (a) compares binaural models and (b) single-channel models for both DEC and Dcor MBSTOI.

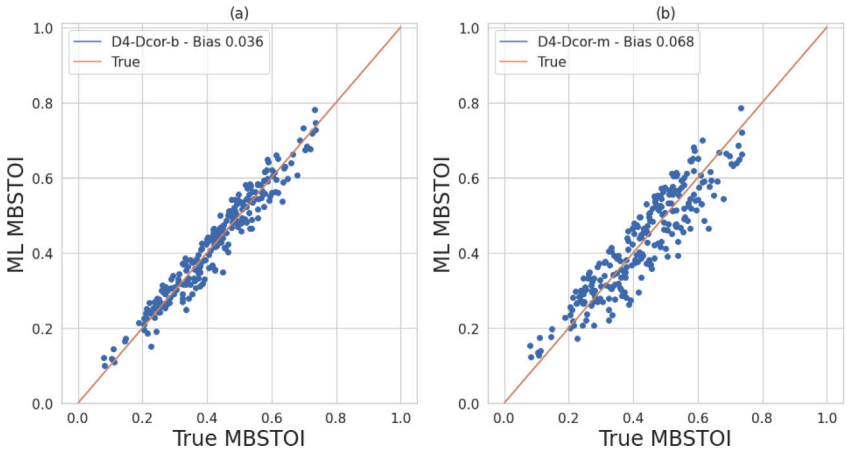

Fig. 11. Scatter plot of Dcor-MBSTOI trained on D4 using (a) binaural reference, $-b$, and (b) single-channel reference, $-m$. Results are presented with bias correction for both the binaural and single-channel metrics.

added reverb (D3 and D4) only brings the models to the level of MBSTOI$_{EC=0}$. This is corroborated by Figs. 6 and 7 where estimated EC parameters are mostly around 0. Notably, no statistical differences can be established between DEC and Dcor single channel across all datasets according to Table 5.

The poor estimation of EC parameters in DEC models from Figs. 6 and 7, discussed in Section 5.1, still led to good DEC-MBSTOI performance in Fig. 10. On the contrary, while Dcor shows good correlation in Fig. 9, it led to poorer performance in Dcor-MBSTOI in Fig. 10. It was shown in Guiraud et al. (2022) that MBSTOI is relatively robust to misestimation of EC parameters hence why DEC has more stable performance. A direct misestimation in MBSTOI correlation coefficients leads to bigger changes, making Dcor-MBSTOI more prone to errors.

It has been shown that while Dcor is more efficient as a machine learning model than DEC, the subsequent DEC-MBSTOI metric is more accurate than Dcor-MBSTOI. This is particularly true when using a binaural reference but less so with single-channel reference.

5.2.2. Bias correction

Due to the construction of MBSTOI, our hybrid DL models output a consistent underestimation of the target value as observed in Fig. 10. To prevent this tendency a bias correction term, called Bias, is calculated. This Bias is then consistently subtracted after the DL-MBSTOI estimation. To include the estimation of this term within the models' construction, it is calculated after training on the Test set. The Bias is chosen to be the mean deviation from MBSTOI defined as

$$\text{Bias} = E\,(\text{x-MBSTOI} - \text{MBSTOI})\,, \tag{1}$$

where E is the arithmetic mean and x-MBSTOI is either DEC-MBSTOI or Dcor-MBSTOI. All following plots showing results on the Validation set now include their model's corresponding Bias correction calculated from the D4 Test set.

Fig. 11 shows scatter plots of Dcor trained on D4 using binaural reference in (a) and single-channel reference in (b) both with their respective bias correction. Strong correlation between true and estimated MBSTOI is observed with a wider spread for the single-channel case. In other words, the Pearson correlation coefficient and the standard deviation are 0.972 and 0.034 with binaural reference and 0.914 and 0.058 with single-channel reference. The bias correction allows to retain the positive association between true and estimated while focusing the data around the true value. Only Dcor-MBSTOI is presented here but similar plots are obtained with DEC-MBSTOI.

5.3. Validation clarity challenge data

Previous results have all been on unseen data from the Validation set. However, this Validation set has been generated in a similar fashion to the Train and Test set. It is then necessary to see how the models perform when confronted with data from other scenarios and collections. To that end, the validation dataset scenes provided by the Clarity challenge has been used (Graetzer et al., 2021). Similarly to our

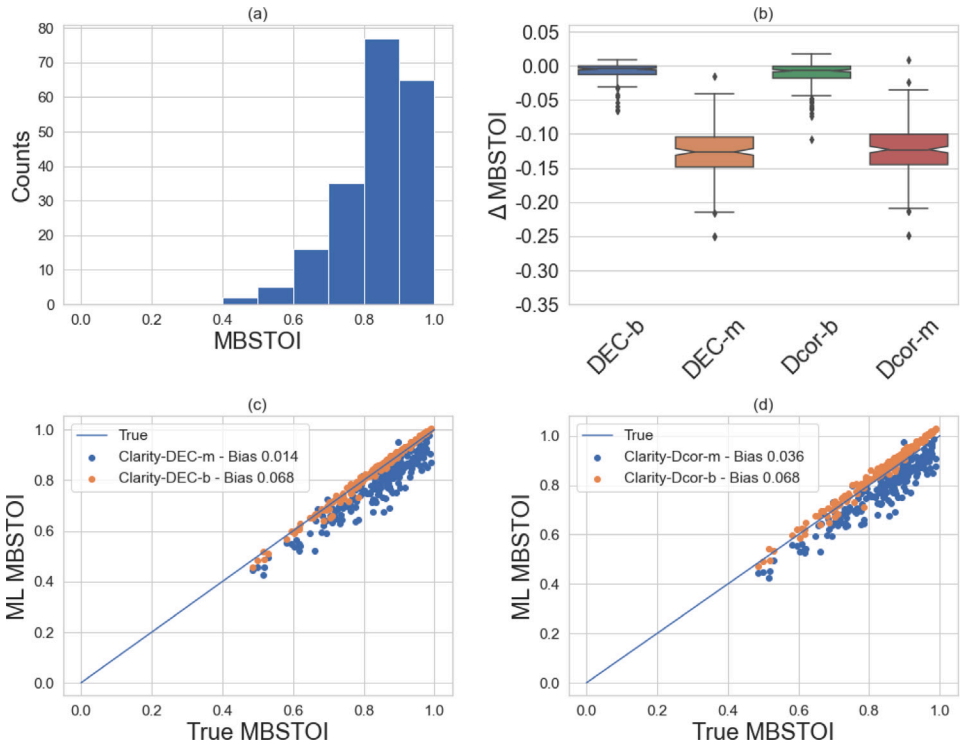

Fig. 12. Clarity challenge dataset analysis. Figure (a) display the MBSTOI distribution of the validation set. Figure (b) shows the difference in MBSTOI estimation with both DEC- and Dcor-MBSTOI, trained on D4 using binaural, $-b$, or single-channel reference, $-m$. Figures (c) and (d) then show the scatter plots of estimated and true MBSTOI.

data, they consist of binaural audio with a localised speech in noise. The results are presented here.

Fig. 12(a) shows the MBSTOI distribution of the Clarity sentences used. It is observed that most sentences have a high MBSTOI. While this does not affect the model, it is here for reference purposes and to show complementary distribution with D4 Validation set as seen in Fig. 3. Fig. 12(b) shows the distribution of ΔMBSTOI across the set using DEC-MBSTOI and Dcor-MBSTOI metrics trained on D4, using both single-channel and binaural reference. Fig. 12(c) and (d) displays the scatter plots of those models after applying the same bias correction used in Fig. 11, calculated on D4 Test set.

As for the simulated datasets, the binaural models perform better as they have a mean absolute deviation of 0.01 and standard deviation less than 0.02, while single-channel reference mean absolute deviation is of 0.12 and standard deviation up to 0.04. Nonetheless, Fig. 12(c) and (d) show good correlation using single-channel reference. The Pearson correlation coefficient is of 0.94, as opposed to more than 0.99 with binaural reference. Overall, use of a binaural reference leads to better performance than single-channel reference models but no statistical difference can be observed in Table 5 between DEC and Dcor on the Clarity dataset.

6. Conclusion and future work

In this work, a hybrid ML-based method to calculate the MBSTOI is presented. The two DEC- and Dcor-MBSTOI metrics replace partially or entirely the EC stage to estimate either the EC parameters or else the correlation coefficients at every time frame and third-octave band. To this end, multiple datasets of increasing difficulty have been created and two variants of each metrics have been tested: the first one using a binaural reference as for MBSTOI, and the other using a single-channel reference.

Presented results show that DEC-MBSTOI with binaural reference can very accurately reproduce MBSTOI with a mean absolute deviation and standard deviation of 0.016 and 0.019 in the most complex scenario with interferer, diffuse noise and reverberation. Dcor-MBSTOI did not perform as well with binaural reference with mean absolute deviation and standard deviation of 0.037 and 0.034 in the same scenario. The DEC DNN model showed poor estimation γ and τ, whereas Dcor model was more accurate in estimating the correlation coefficients. Nonetheless, due to the robustness of MBSTOI in EC parameters variation, DEC-MBSTOI perform better overall.

A decrease in performance is observed for DEC-MBSTOI when using single-channel reference with a mean absolute deviation and standard deviation of 0.068 and 0.056 in the same scenario. Similar performance is achieved with Dcor-MBSTOI using single-channel reference, with a mean absolute deviation and standard deviation of 0.069 and 0.058 in the same scenario. However, it has been shown that single-channel reference model are not statistically different from MBSTOI with a constant estimation of the EC parameters to be 0.

Predictions were improved by correcting for bias (mean deviation), centring metric results around the target value. Similar performance is also observed when using the Clarity challenge dataset indicating that DEC- and Dcor-MBSTOI could be suitable for use in real world scenarios.

By replacing only parts of the MBSTOI estimation by a DL model, hybrid metrics have been created where an originally binaural metric can now be computed using only a single-channel reference. Both hybrid metrics raise the question of how best to integrate ML into an existing tool to improve it, while retaining the original idea. Those metrics extend the range of use of speech intelligibility estimation, for instance to binaural signal processing experiments of a source speech signal.

CRediT authorship contribution statement

Pierre Guiraud: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualisation, Writing – original draft, Writing – review & editing. **Alastair H. Moore:** Conceptualization, Supervision, Writing – review & editing. **Rebecca R. Vos:** Conceptualization, Writing – review & editing. **Patrick A.**

Naylor: Conceptualization, Supervision, Project administration, Funding acquisition, Resources, Writing – review & editing. **Mike Brookes:** Conceptualization, Supervision, Writing – review & editing.

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.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council, United Kingdom [grant number EP/S035842/1].

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., 2016. TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 [cs.DC].

Andersen, A.H., de Haan, J.M., Tan, Z.H., Jensen, J., 2016. Predicting the intelligibility of noisy and nonlinearly processed binaural speech. IEEE/ACM Trans. Audio, Speech, Lang. Process. 24 (11), 1908–1920. http://dx.doi.org/10.1109/TASLP.2016.2588002.

Andersen, A.H., de Haan, J.M., Tan, Z.H., Jensen, J., 2018a. Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (10), 1925–1939. http://dx.doi.org/10.1109/TASLP.2018.2847459.

Andersen, A.H., de Haan, J.M., Tan, Z.H., Jensen, J., 2018b. Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions. Speech Commun. 102, 1–13. http://dx.doi.org/10.1016/j.specom.2018.06.001.

ANSI, 1997. Methods for the Calculation of the Speech Intelligibility Index. ANSI Standard S3.5-1997 (R2007), American National Standards Institute (ANSI).

Beutelmann, R., Brand, T., 2006. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 120, 331–342. http://dx.doi.org/10.1121/1.2202888.

Beutelmann, R., Brand, T., Kollmeier, B., 2010. Revision, extension, and evaluation of a binaural speech intelligibility model. J. Acoust. Soc. Am. 127 (4), 2479–2497. http://dx.doi.org/10.1121/1.3295575.

Bouckaert, R.R., Frank, E., 2004. Evaluating the replicability of significance tests for comparing learning algorithms. In: Proc. Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD).

Cauchi, B., Siedenburg, K., Santos, J.F., Falk, T.H., Doclo, S., Goetze, S., 2019. Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network. IEEE/ACM Trans. Audio, Speech, Lang. Process. 27 (7), 1151–1163. http://dx.doi.org/10.1109/TASLP.2019.2912123.

Chabot-Leclerc, A., MacDonald, E.N., Dau, T., 2016. Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain. J. Acoust. Soc. Am. 140 (1), 192–205. http://dx.doi.org/10.1121/1.4954254.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised claddification learning algorithms. Neural Comput. 10 (7), 1895–1923. http://dx.doi.org/10.1162/089976698300017197.

Durlach, N.I., 1963. Equalization and Cancellation Theory of Binaural Masking-Level Differences. J. Acoust. Soc. Am. 35 (8), 1206–1218. http://dx.doi.org/10.1121/1.1918675.

French, N.R., Steinberg, J.C., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19 (1), 90–119. http://dx.doi.org/10.1121/1.1916407.

Graetzer, S., Barker, T., Cox, T.J., Akeroyd, M., Culling, J.F., Naylor, G., Porter, E., Viveros Muñoz, R., 2021. Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing. In: Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH). Brno, Czech Republic, http://dx.doi.org/10.21437/Interspeech.2021-1574.

Grimm, G., Luberadzka, J., Hohmann, V., 2019. A toolbox for rendering virtual acoustic environments in the context of audiology. Acta Acust. United Acust. 105 (3), 566–578. http://dx.doi.org/10.3813/AAA.919337.

Guiraud, P., Moore, A.H., Vos, R.R., Naylor, P.A., Brookes, M., 2022. Machine learning for parameter estimation in the MBSTOI binaural intelligibility metric. In: Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC).

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio, Speech, Lang. Process. 16 (1), 229–238. http://dx.doi.org/10.1109/TASL.2007.911054.

Hu, G., Wang, D.L., 2010. A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans. Audio, Speech, Lang. Process. 18 (8), 2067–2079. http://dx.doi.org/10.1109/TASL.2010.2041110.

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 (11), 2009–2022. http://dx.doi.org/10.1109/TASLP.2016.2585878.

Jørgensen, S., Dau, T., 2011. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. J. Acoust. Soc. Am. 130 (3), 1475–1487. http://dx.doi.org/10.1121/1.3621502.

Jørgensen, S., Ewert, S.D., Dau, T., 2013. A multi-resolution envelope-power based model for speech intelligibility. J. Acoust. Soc. Am. 134 (1), 436–446. http://dx.doi.org/10.1121/1.4807563.

Kates, J.M., Arehart, K.H., 2014. The hearing-aid speech perception index (HASPI). Speech Commun. 65, 75–93. http://dx.doi.org/10.1016/j.specom.2014.06.002.

Kates, J.M., Arehart, K.H., 2021. The hearing-aid speech perception index (HASPI) version 2. Speech Commun. 131, 35–46. http://dx.doi.org/10.1016/j.specom.2020.05.001.

Kim, D.S., Tarraf, A., 2007. ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality. Bell Labs Tech. J. 12, 221–236. http://dx.doi.org/10.1002/bltj.20228.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proc. Intl. Conf. on Learning Representations (ICLR). http://dx.doi.org/10.48550/arXiv.1412.6980.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc, pp. 8024–8035.

Pedersen, M.B., Andersen, A.H., Jensen, S.H., Jensen, J., 2020. A Neural Network for Monaural Intrusive Speech Intelligibility Prediction. In: Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP). http://dx.doi.org/10.1109/ICASSP40776.2020.9052949.

Rosbach, J., Rottges, S., Hauth, C.F., Brand, T., Meyer, B.T., 2021. Non-Intrusive Binaural Prediction of Speech Intelligibility Based on Phoneme Classification. In: Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP). Institute of Electrical and Electronics Engineers (IEEE), pp. 396–400. http://dx.doi.org/10.1109/icassp39728.2021.9413874.

Rothauser, E.H., Chapman, W.D., Guttman, N., Hecker, M.H.L., Nordby, K.S., Silbiger, H.R., Urbanek, G.E., Weinstock, M., 1969. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 17 (3), 225–246. http://dx.doi.org/10.1109/TAU.1969.1162058.

Schomer, P.D., Swenson, G.W., 2002. 40 - Electroacoustics. In: Middleton, W.M., Van Valkenburg, M.E. (Eds.), Reference Data for Engineers (Ninth Edition), Ninth Edition Newnes, Woburn, 40–1–40–28. http://dx.doi.org/10.1016/B978-075067291-7/50042-X.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio, Speech, Lang. Process. 19 (7), 2125–2136. http://dx.doi.org/10.1109/TASL.2011.2114881.

Vanwinckelen, G., Blockeel, H., 2012. On estimating model accuracy with repeated cross-validation. In: Proc. Belgian-Dutch Conf. on Machine Learning (BeneLearn2012).

Yao, Y., Rosasco, L., Caponnetto, A., 2007. On early stopping in gradient descent learning. Constr. Approx. 26, 289–315. http://dx.doi.org/10.1007/s00365-006-0663-2.

Zezario, R.E., Fu, S.-W., Fuh, C.-S., Tsao, Y., Wang, H.-M., 2020. STOI-Net: A Deep Learning based Non-Intrusive Intelligibility Assessment Model. In: Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC).

Zhang, T., Yu, B., 2005. Boosting with early stopping: Convergence and consistency. Ann. Statist. 33 (4), 1538–1579. http://dx.doi.org/10.1214/009053605000000255.