This is a repository copy of *Non-intrusive speech intelligibility estimated by metric prediction for hearing impaired individuals for the clarity prediction challenge 1*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/189718/

Version: Published Version

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Non-intrusive Speech Intelligibility Estimated By Metric Prediction for Hearing Impaired Individuals for the Clarity Prediction Challenge 1

George Close [ID] *, Samuel Hollands* [ID] *, Thomas Hain* [ID] *, and Stefan Goetze* [ID]

UKRI CDT for Speech and Language Technologies and their Applications,
Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

{glclose1,shollands1,t.hain,s.goetze}@sheffield.ac.uk

Abstract

This paper proposes neural models to predict Speech Intelligibility (SI),both by prediction of established SI metrics and of human speech recognition (HSR) on the 1st Clarity Prediction Challenge. Both intrusive and non-intrusive predictors for intrusive SI metrics are trained, then fine-tuned on the HSR ground truth. Results are reported on a number of SI metrics, and the model choice for the Clarity challenge submission is explained. Additionally, the relationship between the SI scores in the data and commonly used signal processing metrics which approximate SI are analysed, and some issues emerging from this relationship discussed. It is found that intrusive neural predictors of SI metrics when fine-tuned on the true HSR scores outperform the non neural challenge baseline.

1. Introduction

In the United Kingdom (UK) 1 in 5, or just over 12 million people, experience Hearing Loss (HL) of greater than 25 decibels hearing level (dBHL) [1]. By 2035 this will rise to 14.2 million [1] and with age correlating with an individual's likelihood for developing impaired hearing this statistic is going to inflate dramatically. By 2050 we will have observed a near doubling of the global population aged older than 60 going from just 12% in 2015, to making up 22% of the world's population by 2050 [2]; a reality that has large consequences for all medical conditions which increase in likelihood with age.

Whilst it may seem intuitive to imagine that hearing aids are somewhat of a solved problem given their supposed ubiquitousness in society, the reality is far from that case. 80% of adults aged 55-74 who would benefit from wearing a hearing aid do not use one, this accounts for approximately 32% of the entire 55-74 population [3]. Many reasons are given including poor fit, side effects such as rashes, technical difficulties replacing cells, and a dislike of the aesthetic [3]. However studies repeatedly demonstrate that the main issue prompting users to not use their hearing aids is poor performance, often in environments with background noise [3, 4, 5, 6, 7, 8] leading to decreased HSR [9]. Untreated HL can result in many harmful outcomes such as a loss of environmental awareness leading to accidents, a general worsening of quality of life, and even greater risks for developing depression [4] as well as impacts on neurological conditions [10]. In the context of the Clarity Prediction Challenge 1 (CPC1) [11] Speech Intelligibility (SI) is defined as the percentage of words that a listener correctly identifies after listening to a sequence of words.

SI prediction metrics are either *intrusive*, i.e. rely on access to the clean reference signal, or *non-intrusive*, i.e. rather than facilitating a comparative function non-intrusive metrics analyse only the degraded signal under test to identify key areas of potential distortion [12]. There are 3 key domains of non-intrusive SI; feature-based approaches using key acoustic features and potentially other linguistic information for prediction, statistical data-driven methods such as machine learning, and neurophysiological measures that integrate neuroimaging or oculometric techniques [13]. This paper aims to use both non-intrusive and intrusive methods for predicting intrusive SI metrics as outlined in Section 2.

2. Speech Intelligibility Metrics

This paper explores the viability of a neural network to predict intrusive metrics for SI using non-intrusive and intrusive input audio. 3 intrusive SI metrics listed here in increasing levels of complexity of computation are investigated. All metrics return a number between 0% and 100% representing the percentage of words correctly identified.

The Short-Time Objective Intelligibility (STOI) [14] is a commonly used monoaural metric for the assessment of SI. It works by computing an average of the correlation between one-third-octave filter-bank representations of the clean and degraded speech signals. It has been found to correlate well with human intelligibility in normal hearing individuals [15, 16, 17].

The Modified Binaural Short-Time Objective Intelligibility (MBSTOI) [18] is a variant of STOI which takes into account binaural degraded and reference signals. The score additionally includes internal simulation of the 'better ear effect' wherein the channel with the highest correlation for that block of processing is used to compute the final score.

The Hearing-Aid Speech Perception Index (HASPI) [19] is designed specifically to assess intelligibility in people with HL. In addition to a degraded and a reference signal it also takes an audiogram representation of the HL in a given ear into account, and incorporates a HL simulation as part of the computation of the score. It additionally incorporates an ensemble of neural networks fitted to real human intelligibility as part of the score calculation.

2.1. CPC1 Score Distributions

It is worth noting that, compared to other SI corpora including ground truth HSR scores, data used in this challenge contains grammatically intact sentences. This has been demonstrated to yield lower accuracy for diagnosis than a bag of words approach [20] in optometry or matrix tests [21] in SI assessments. Most participant in the CPC1 data have moderate to severe high-frequency HL as Figure 1 shows in terms of box plots
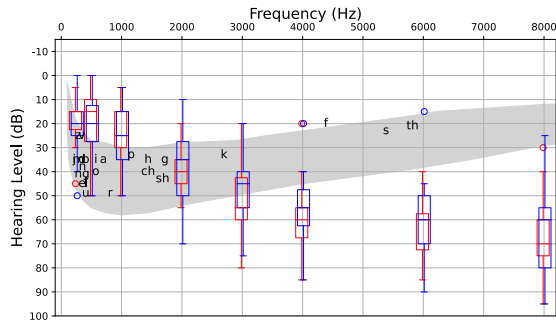
Figure 1: *Audiogram data for left (red) and right (blue) ear of in CPC1 listeners and the frequency characteristics of speech phones*

for audiograms of left and right ear in red and blue colour, respectively. The grey shaded area in Figure 1 shows typical level range and frequency distribution as well as positions of individual speech fragments according to their principal frequency content [22]. This indicates that phones such as /k/, /f/, /s/ or /th/, but also others, are usually not perceived by the participants since their (central) energy lies below the hearing threshold. It is therefore important to consider whether the purpose of a hearing aid algorithm evaluation is designed to maximise language comprehension given an individual's cognitive language processing abilities, or to maximise signal enhancement for language carrying signals agnostic of cognitive ability. This creates a problem for evaluation metrics as such metrics fall firmly within the latter, but querying an individual to repeat a grammatical sentence back falls firmly in the former definition. Beyond grammar one might even remark upon the likely less impactful entropy of phones themselves and how an individual guessing a word is fundamentally depending on known rules of language and thus are still to a degree dependent on cognitive ability. Future experiments looking to test the impact of linguistic knowledge may wish to try comparing the performance of individuals with no background in English to determine to what degree these metrics are truly agnostic of cognitive and linguistic ability, if the objective is to maximise the enhancement ability of a hearing aid algorithm agnostic of human language comprehension.
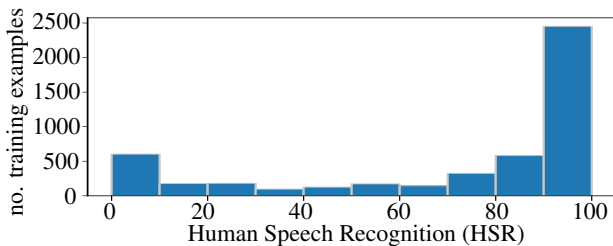

Figure 2: *Histogram showing distribution of ground truth correctness Q_h, i.e. HSR in CPC1 Training set*

Figure 2 shows the distribution of the ground truth intelligibility scores in the CPC1 training data. This shows that the largest class is a HSR of 100 where the listeners were able to perfectly identify the sequence of words they listened to. The next largest class is an HSR of 0 where the listener was unable

to identify any of the words; either guessing 0 words correctly or not making an attempt.

The Spearman r and Pearson ρ correlations between the SI metrics and the ground truth HSR are presented in Table 1 and the relationships are visualised in Figure 3.

Metric	STOI	MBSTOI	HASPI
r	0.65	0.61	0.34
ρ	0.57	0.54	0.31

Table 1: *Spearman r and Pearson ρ Correlation between SI metrics and HSR*

These both show that, while correlation is low for all three metrics, STOI and MBSTOI correlate somewhat more strongly with the data compared with HASPI - this is interesting, especially given that HASPI is the one metric of the 3 which has explicit access to the audiogram information. One possible explanation is that STOI and MBSTOI are computed using $\hat{x}$ while HASPI uses x as it contains its own internal HL simulation; it is possible that this internal model produces outputs which differ greatly from that of the baseline system.

3. Neural Intelligibility Prediction

Inspired by recent works [23, 24, 25] which use a neural network to mimic the performance of an intrusive metric for speech quality and intelligibility, this contribution uses a similar network structure to predict the metric score that will be assigned to the input audio. Note that here networks that are provided with representations of both the degraded and reference signal (intrusive) and also with those that are only provided with the degraded (non-intruisve) are investigated.
The focus is on a metric prediction objective over simply using the ground truth 'correctness' information in the training data as this was found to be distributed in a way that was difficult for our non-intrusive models to find any discernible patterns in. Intuition is that if these metrics have been found to correlate with human intelligibility, then non-intrusive predictors of said metrics should also. Additionally, the performance of each of our non-intrusive metric predictors after being fine-tuned on the ground truth intelligibility is reported.
Figure 4 provides a generalised overview demonstrating the training of such a neural network. Here, noisy audio x is generated by a Speech In Noise (SPIN) generator and processed by a hearing aid (HA) simulation then a HL simulation which both take a representation of the specific listener's HL as input. This takes the form of an audiogram pair $\{\mathbf{a}^l, \mathbf{a}^r\}$ which represent the specific characteristics of their HL for the left and right ears respectively. Details on the HL model used in the CPC1 baseline can be found in [26]. The output of this $\hat{x}$ is input to a SI prediction model, along with the clean reference audio s. The output of this prediction model $\hat{Q}$ is compared to the true value of the SI Q i.e the HSR, and the model is updated.

3.1. Feature Extraction

The same feature extraction as described in [25] is used here with the discrete time domain input audio being transformed to normalised log features. Note that in the following $\mathbf{X}_f^l$, $\mathbf{X}_f^r$ denotes the feature representation of the hearing aid output while $\hat{\mathbf{X}}_f^l$, $\hat{\mathbf{X}}_f^r$ is the feature representation of the hearing aid output x with the baseline HL applied $\hat{x}$.
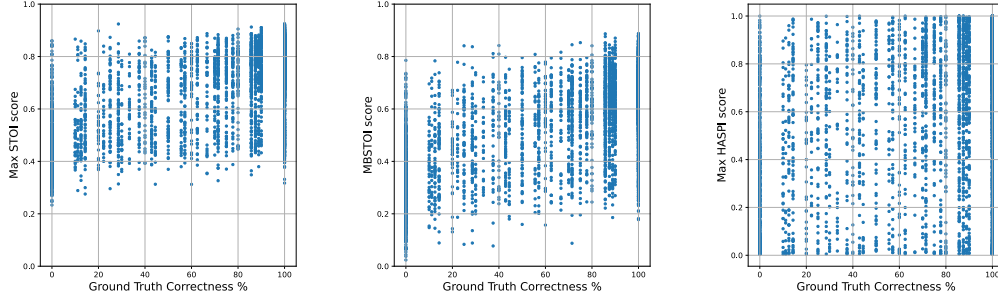
Figure 3: *SI metrics versus ground truth HSR in Clarity Prediction Challenge Training Set*
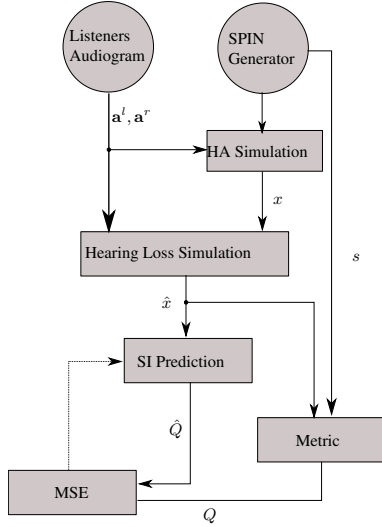

Figure 4: *Diagram of general non-intrusive SI metric prediction training*

3.2. Model Structure for Non-Intrusive Prediction

For each of the 3 metrics investigated, the same basic model structure is adapted for the specific requirements of the metric. The basic structure is based on that of the discriminator network in [24] - 4 2D convolutional layers with 15 filters of a kernel size of $(5, 5)$. To account for the variable length of input data, a global 2D average pooling layer is placed immediately after the input, fixing the feature representation at 15 dimensions. After the convolutional layers, a mean is taken over the 2nd and 3rd dimensions, and this representation is fed into 3 sequential linear layers, with 50, 10, and 1 output neuron(s) respectively. The first 2 of these layers have a LeakyReLU activation while the final layer has no activation. For STOI, the score for each channel of the HA output audio $\hat{x}$ is predicted separately, with the input to the prediction network being the feature space representation of the given channel $\hat{\mathbf{X}}_f^c$ where c is a channel index. As such, the input dimension to the average pooling and first 2D convolutional layer is set to 1. For MBSTOI, the score is predicted for the HA output stereo audio together, with the input to the network being the feature space representations of both channels $\{\hat{\mathbf{X}}_f^l, \hat{\mathbf{X}}_f^r\}$. The input dimension of the average pooling layer and the initial convolutional layer is 2 to account for these stacked channel representations. Finally for HASPI which like STOI is defined per audio channel, the $\mathbf{X}_f^l$ and $\mathbf{X}_f^r$ representation of the audio, but also use

$\mathbf{a}^l$ and $\mathbf{a}^r$ the audiogram representations of the listener's HL is used as input. This 6 element representation is passed through a linear layer with 10 output neurons then another with 50; this representation is then concatenated along the feature dimension with the representation of the audio of the same size. This 100 element representation is then fed through a further 3 linear layers with 50, 10, and 1 output node(s) respectively, all but the last layer having a LeakyReLU activation. Additionally, we train a model with the same structure as that for the HASPI prediction described above, using $\{\hat{\mathbf{X}}_f^l, \hat{\mathbf{X}}_f^r\}$ as input and train it to predict the ground truth Correctness scores in the training data.

3.3. Model Struture for Intrusive Prediction

We additionally train and fine-tune intrusive versions of the metric prediction models. These are similar to those above except we also input the clean reference features $\mathbf{S}_f^c$ to the model. For the STOI and HASPI prediction models we stack the clean and degraded features per channel, $\{\mathbf{S}_f^l, \hat{\mathbf{X}}_f^l\}$, $\{\mathbf{S}_f^r, \hat{\mathbf{X}}_f^r\}$ for STOI and $\{\mathbf{S}_f^l, \mathbf{X}_f^l\}, \{\mathbf{S}_f^r, \mathbf{X}_f^r\}$ along with $\mathbf{a}^l$, $\mathbf{a}^r$ for HASPI. For MBSTOI we use both channels of the clean and degraded features, $\{\hat{\mathbf{X}}_f^l, \mathbf{S}_f^l, \hat{\mathbf{X}}_f^r, \mathbf{S}_f^r\}$.

4. Experiments

4.1. Tools and Software

Experiments are implemented via modifications to the challenge baseline system, replacing the simple fitting model with the neural models described above using PyTorch [27]. The SpeechBrain [28] framework is used for audio loading and dataloader creation. Existing Python and MATLAB implementations are used for STOI [1], MBSTOI (taken from CPC1 baseline) and HASPI [2]. All of the models are relatively low cost, and can be run on a CPU in a reasonable amount of time.

4.2. Data Description

Audio data provided by the CPC1 is used for the hearing aid outputs x, the hearing aid outputs processed by the baseline HL simulation $\hat{x}$ and the anechoic clean reference signal s, accompanied by ground truth correctness scores Q_h and listeners' audiograms $\{\mathbf{a}^l, \mathbf{a}^r\}$ for left and right ear, respectively. In total the challenge corpus provides 4863 training examples expressed as combinations of 'scenes' (s, x), listener HL characteristics $(\mathbf{a}^l, \mathbf{a}^r)$, HL simulations $\hat{x}$ and correctness scores Q_h. The spoken sentence are taken from the Clarity speech corpus

---

[1]https://github.com/mpariente/pystoi
[2]https://claritychallenge.github.io/clarity_CPC1_doc/docs/cpc1_faq

[29].

4.3. Experiment Setup

We pre-compute the STOI, MBSTOI, and HASPI scores for the entire train set. We then train models as described above, to reproduce the score. The feature extraction use a Short Time Fourier Transform (STFT) with a window length of 20ms, a hop length of 10ms and an FFT size of 1024. The hearing aid outputs x have a sampling rate of 32kHz, while the hearing aid outputs with the baseline HL simulation applied $\hat{x}$ have a sampling rate of 44kHz. Following on from the baseline system we train with a 5 fold validation technique, partitioning the folds on the scene ID. We use the Adam [30] Optimiser with a learning rate of 0.001 for all models. All models are trained with a batch size of 1 with the exception of the model that directly predicts Correctness which uses a batch size of 20. The metric prediction models are additionally fine-tuned using the ground truth HSR 'Correctness' (intelligibility) scores; in the case of the the metrics that are defined per channel (STOI and HASPI) we use the channel that returned the highest predicted score between the 2, as a simplified simulation of the 'better ear effect'. This fine-tuning process consists of exposing the model to the entire training set in the same way as in the pre-training, but having it's outputs compared to the ground truth rather than the metric. We use this same technique to evaluate the performance of these models.

5. Results

Table 2 shows the results for non-intrusive prediction over the entire training set for the challenge. The upper half shows the Root Mean Square Error (RMSE) between model output and ground truth 'correctness' values, i.e. HSR. The lower half shows the RMSE between target metric and prediction of the model. r and ρ are the Spearman and Pearson Correlations, respectively.

In terms of prediction error, the model showing best non-

Table 2: *Non Intrusive Performance on the Clarity Prediction Challenge Training Set*

Model Objective	Correctness Error	r	ρ
STOI	35.63	0.30	0.21
STOI (fine)	34.55	0.32	0.25
MBSTOI	39.30	0.26	0.18
MBSTOI (fine)	34.72	0.32	0.23
HASPI	38.80	0.23	0.22
HASPI (fine)	**31.55**	**0.53**	**0.46**
Correctness	33.44	0.45	0.42
	Prediction Error	r	ρ
STOI	**13.88**	0.43	0.3
STOI (fine)	16.44	0.43	0.3
MBSTOI	15.50	0.44	0.33
MBSTOI (fine)	21.81	0.47	0.32
HASPI	25.10	**0.59**	**0.59**
HASPI (fine)	37.09	0.29	0.29

intrusive target metric prediction is the STOI prediction model, while the the HASPI model shows lowest performance. This is likely because the calculation of STOI is considerably simpler than that for HASPI. As expected, fine-tuning to the ground truth correctness increases prediction error while decreasing correctness error for all models.

Table 3: *Intrusive Performance on the Clarity Prediction Challenge Training Set*

Model Objective	Correctness Error	r	ρ
baseline	*28.5*	*0.62*	*0.54*
STOI	32.45	0.58	0.52
STOI (fine)	27.59	0.66	0.56
MBSTOI	29.67	0.65	0.54
MBSTOI (fine)	**27.20**	**0.67**	**0.58**
HASPI	41.04	0.27	0.25
HASPI (fine)	29.67	0.65	0.54
Correctness	35.62	0.31	0.27
	Prediction Error	r	ρ
STOI	**9.05**	**0.86**	**0.83**
STOI (fine)	16.24	0.75	0.70
MBSTOI	10.79	0.79	0.80
MBSTOI (fine)	22.64	0.73	0.7
HASPI	23.06	0.68	0.68
HASPI (fine)	29.11	0.43	0.43

Table 4: *Non Intrusive Performance on the Clarity Prediction Challenge Test Set*

Model Objective	Correctness Error	r	ρ
HASPI (fine)	**31.99**	**0.43**	**0.50**
Correctness	33.42	0.42	0.39

Best model in terms of prediction of ground truth correctness is the fine-tuned HASPI predictor. This is interesting given that HASPI itself has the lowest correlation with the ground truth correctness in the data - it is possible that access to the audiogram information is what enables this. The slight performance improvement versus the model that was only trained to predict the correctness shows that the HASPI objective pre-training did improve performance.

Table 3 shows the results of the intrusive prediction models, along with that of the challenge baseline system. The prediction error results follow the same pattern as those of the non-intrusive models, but with lower overall error rates and significantly higher correlations.

Both the fine-tuned STOI and MBSTOI models slightly outperformed the baseline system in terms of correctness error and correlations.Interestingly, of the two models that directly predict the Correctness values Q, the non-intrusive model slightly outperforms the intrusive one.

Table 4 shows the performance on the test set of the two non-intrusive models submitted to the challenge. The pretrained HASPI model performs slightly better overall compared to the direct Correctness model.

6. Conclusion

Of the models trained, it was found that intrusive models outperform non intrusive models for both metric prediction and for real intelligibility prediction. An intrusive neural model outperforms the intrusive baseline system for the challenge. Furthermore, pre-training models to predict an intelligibility metric, and then fine-tuning on the true intelligibility improves performance. Additionally, the relationship between the real intelligibility scores in the data and signal processing based intrusive metrics was examined, and it was found that these are only weakly correlated.

7. References

[1] N. Park, "Population estimates for the uk, england and wales, scotland and northern ireland, provisional: mid-2019," *Hampshire: Office for National Statistics*, 2020.

[2] World Health Organization, *World report on ageing and health*. World Health Organization, 2015.

[3] A. McCormack and H. Fortnum, "Why do people fitted with hearing aids not wear them?" *International journal of audiology*, vol. 52, no. 5, pp. 360–368, 2013.

[4] S. Arlinger, "Negative consequences of uncorrected hearing loss-a review," *International journal of audiology*, vol. 42, pp. 2S17–2S20, 2003.

[5] H. Arksey and L. O'Malley, "Scoping studies: towards a methodological framework," *International journal of social research methodology*, vol. 8, no. 1, pp. 19–32, 2005.

[6] W. Chien and F. R. Lin, "Prevalence of hearing aid use among older adults in the united states," *Archives of internal medicine*, vol. 172, no. 3, pp. 292–293, 2012.

[7] J. Cohen-Mansfield and J. W. Taylor, "Hearing aid use in nursing homes, part 2: Barriers to effective utilization of hearing aids," *Journal of the American Medical Directors Association*, vol. 5, no. 5, pp. 289–296, 2004.

[8] K. Davis, N. Drey, and D. Gould, "What are scoping studies? a review of the nursing literature," *International journal of nursing studies*, vol. 46, no. 10, pp. 1386–1400, 2009.

[9] C. Spille, B. Kollmeier, and B. T. Meyer, "Comparing human and automatic speech recognition in simple and complex acoustic scenes," *Computer Speech & Language*, vol. 52, pp. 123–140, 2018.

[10] F. R. Lin, E. J. Metter, R. J. O'Brien, S. M. Resnick, A. B. Zonderman, and L. Ferrucci, "Hearing loss and incident dementia," *Archives of neurology*, vol. 68, no. 2, pp. 214–220, 2011.

[11] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proceedings of the 23nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2022)*, Incheon, Korea, Sep. 2022.

[12] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[13] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, p. 103204, 2022.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[15] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[16] S. Goetze, A. Warzybok, I. Kodrasi, J. Jungmann, B. Cauchi, J. Rennies, E. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A Study on Speech Quality and Speech Intelligibility Measures for Quality Assessment of Single-Channel Dereverberation Algorithms," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC 2014)*, Antibes, France, Sep. 2014.

[17] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors," in *2020 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, October 2020, pp. 6493–6497.

[18] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[19] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[20] M. MacKeben, U. K. Nair, L. L. Walker, and D. C. Fletcher, "Random word recognition chart helps scotoma assessment in low vision," *Optometry and Vision Science*, vol. 92, no. 4, p. 421, 2015.

[21] B. Kollmeier, A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, and K. C. Wagener, "The multilingual matrix test: Principles, applications, and comparison across languages: A review," *International Journal of Audiology*, vol. 54, no. sup2, pp. 3–16, 2015, pMID: 26383182.

[22] J. Rennies, S. Goetze, and J.-E. Appell, "Personalized Acoustic Interfaces for Human-Computer Interaction," in *Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications*, M. Ziefle and C.Röcker, Eds. IGI Global, 2011, ch. 8, pp. 180–207.

[23] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," 2018.

[24] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," 2021.

[25] G. Close, T. Hain, and S. Goetze, "MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data," in *Submitted to EUSIPCO 2022*, Belgrade, Serbia, Aug. 2022, Online: https://arxiv.org/abs/2203.12369.

[26] Y. Nejime and B. C. J. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 603–615, 1997.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[28] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[29] S. Graetzer, M. A. Akeroyd, J. Barker, T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. Viveros-Muñoz, "Dataset of british english speech recordings for psychoacoustics and speech processing research: The clarity speech corpus," *Data in Brief*, vol. 41, p. 107951, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340922001627

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.