Non-intrusive prediction of speech intelligibility for the first Clarity Prediction Challenge (CPC1)

Alex F. McKinney[1], *Benjamin Cauchi*[2]

[1]Department of Computer Science of Durham University, United Kingdom
[2]OFFIS e.V. Institute for Information Technology, Oldenburg, Germany
alexander.f.mckinney@durham.ac.uk, benjamin.cauchi@offis.de

Abstract

Most existing speech intelligibility measures are either designed for single-channel applications – hence unsuited to evaluate hearing aid algorithms –or intrusive, applicable only in simulated scenarios in which the clean signal is available. Non-intrusive speech intelligibility measures able to reliably predict speech intelligibility without knowledge of the clean signal are urgently needed. This paper proposes a non-intrusive measure that predicts speech intelligibility using only the processed signals and audiogram of the listener as input. The proposed measure relies on three steps, namely a hearing-loss model, a feature extractor and a predicting function. The hearing loss model uses the target signal and the listener's audiogram as input while the feature extractor and the predicting function are trained on processed signals labeled in terms of speech intelligibility during a listening test. The evaluation is conducted using cross-validation on both tracks of the first Clarity Prediction Challenge (CPC1).

Index Terms: non-intrusive speech intelligibility prediction; self-supervised learning; contrastive predictive coding

1. Introduction

The number of people suffering from hearing loss is rapidly increasing and despite the progress in hearing aid technology, the problem of hearing aid processing of speech-in-noise remains challenging. One of the many aspects to be addressed in order to solve this issue, is the improvement of the SI measures used to evaluate speech enhancement algorithms. SI represents the ability of listeners to understand speech from signals degraded by noise, reverberation or even processing artefacts. It is often reported using the speech reception threshold (SRT) measured during listening tests [1]. Though typically considered as the gold standard of SI measurements, these tests are costly, time-consuming and often not feasible, e.g., when online estimation of SI is necessary. Consequently, many signal-based measures have been developed. These measures aim at estimating SI without the need for listening tests and can be broadly categorized as being either intrusive or non-intrusive [2]. Intrusive measures are computed using both a clean reference signal and a test signal as input, whereas non-intrusive measures can be computed from the test signal alone. Additionally, SI in signals processed for hearing aid applications largely depends on the presence of binaural cues [3] and measures should be developed for this use case. A reliable non-intrusive SI measure applicable to binaural signals would facilitate the evaluation of binaural speech enhancement algorithms in realistic settings and allow for a better automatic selection of hearing aids parameters.

Most signal-based measures of SI are however designed to be applied only to single-channel signals. Examples of intrusive single-channel SI measures include the articulation index [4], the speech transmission index (STI) [5], the speech intelligibility index (SII) [6], the short-time objective intelligibility (STOI) [7] and mutual-information-based techniques, such as the algorithm proposed in [8]. Several non-intrusive single-channel SI measures have been designed as extensions of the STOI [9, 10], relying on estimating the amplitude envelope of the clean speech from the input signal. Others, such as the speech-to-reverberation modulation energy ratio (SRMR) [11] and its extension the normalized SRMR ($\mathrm{SRMR}_{\mathrm{norm}}$) [12] apply a predicting function on perceptually motivated features extracted from the target signal. SI measures that have been proposed for binaural scenarios include the use combination of equalization-cancellation (EC) models [13] with the SII [14, 15]. The binaural STOI (BSTOI), later refined into the deterministic BSTOI (DBSTOI), uses an EC model to combine both channels of the binaural signal into a single-channel signal used as input to the STOI measure [16]. Both BSTOI and DBSTOI are intrusive.

More recently proposed SI measures rely on the progress in machine learning techniques. This can entail the use of an automatic speech recognizer (ASR), as proposed in [17, 18]. Aiming at non-intrusive prediction, the method in [19], applies the binaural preprocessing stage from [20] to process the binaural signal before using it as input to the ASR. The SI is afterwards predicted by applying mapping between the mean temporal distance (MTD) – a representation of the ASR error [21] – and the SRT. Most machine learning based approaches do not rely on an ASR but rather on a set of features input to a deep neural network. This is the case, for example, in [22], where a neural network predicts SI from a sequence of spectral features, in [23], where both short- and long-term features are input to a classification and regression tree or in [24] , where STOI like features are input to a convolutional neural network [24]. We recently proposed to predict SI from binaural signals by using features computed as a latent representation of the signal as input to a deep learning based SI predictor [25]. These features are computed using a combination of contrastive predictive coding (CPC) [26] and vector quantization (VQ) [27] methods and referred to as VQ-CPC features.

The use of machine learning for SI has however often been burdened by the lack of large datasets of binaural signals labeled in terms of SI. Thanks to the development of the first Clarity Prediction Challenge (CPC1) [28], such dataset is now available to develop and compare SI measures. Taking advantage of

This work was conducted as part of the RISE exchange program funded by the German Academic Exchange Service (DAAD) and supported by the project Augmented Auditory Intelligence (A2I) funded by the German Ministry of Education and Research (BMBF) under grant number 16SV8594.

this opportunity, the work presented in this report has two goals. First it aims to confirm the suitability of VQ-CPC features for SI prediction from binaural signals. Second, it aims at developing a reliable non-intrusive SI measure that could be used in hearing aids applications. For this purpose, the VQ-CPC features are computed from signals pre-processed using an hearing-loss model before being input to a predicting function that improves on the one that we originally used in [25].

The remainder of this report is structured as follows. The proposed non-intrusive SI measure is described in Section 2. The experiments and considered benchmark, based on the CPC1 dataset, are described in Section 3. The results in terms of root mean-squared error (RMSE) are presented in Section 4 and Section 5 concludes the report.

2. Proposed approach

The non-intrusive SI measure that is proposed in this paper is intended to be designed to evaluate the speech intelligibility that a listener with hearing loss would experience from noisy reverberant signal processed through hearing aid processing algorithms. The measure is computed from the audiogram of the target listener and the processed binaural signal $y_m(n)$, where n and $m \in [0, 1]$ denote the sample and channel index, respectively. This computation is done in three steps that are presented in the following subsections.

2.1. Hearing loss model

SI is largely dependent on the type and severity of the hearing loss of the target listener. In order to take this into account, a hearing loss simulator using the Moore, Stone, Baer and Glasberg (MSBG) hearing loss model is used. This model is based on the work of the Cambridge Auditory Group [29, 30, 31, 32]. The implementation provided with the software of the CPC1 baseline [33] is used in this paper. The signal $y_m(n)$ is processed in the gammatone filterbank domain to simulate the four main aspects of hearing loss, namely the raised auditory thresholds, the reduced dynamic range and the lower temporal and frequency resolution. The audiogram of the target listener is used to attenuate the signal in each frequency band according to their hearing loss. The loss in temporal and frequency resolution is modelled through frequency smearing whose amount is dependant on the severity of the listener's hearing loss as described in [28]. The application of this hearing loss model is the only part of the proposed non-intrusive measure that is listener dependent. The output of this stage is a two channel signal $x_m(n)$ from which features are extracted.

2.2. Feature extraction

VQ-CPC features are computed from the two-channel signal $x_m(n)$ using the approach that we recently proposed in [25].

The microphone signal is divided into $T = \lceil N/H \rceil$ overlapping frames of length W, where H denotes the hop length. The samples in each t^{th} frame are used to construct a vector of length $2 \cdot W$:

$$\boldsymbol{x}_t = \left[x_0(tH), \, \dots \, , x_1(tH + W - 1)\right]^{\mathsf{T}} \tag{1}$$

resulting in the time-ordered sequence of T vectors:

$$\boldsymbol{x} = \left\{\boldsymbol{x}_0, \, \boldsymbol{x}_1, \, \dots \, , \boldsymbol{x}_{T-1}\right\}. \tag{2}$$

The feature computation results in the sequence:

$$\boldsymbol{c} = \left\{\boldsymbol{c}_0, \, \boldsymbol{c}_1, \, \dots \, , \boldsymbol{c}_{T-1}\right\}, \tag{3}$$

where $\boldsymbol{c}_t$ denotes the vector of length K feature coefficients extracted from the t^{th} frame. The feature extraction is trained and learns to extract sequences $\boldsymbol{c}$ that maximise the mutual information between the input and output sequences:

$$I(\boldsymbol{x}; \boldsymbol{c}) = \sum_{\boldsymbol{x},\boldsymbol{c}} p\left(\boldsymbol{x}, \boldsymbol{c}\right) \log\left(\frac{p\left(\boldsymbol{x}|\boldsymbol{c}\right)}{p\left(\boldsymbol{x}\right)}\right). \tag{4}$$

To do so, VQ and CPC methods are used to compute the sequence $\boldsymbol{c}$ as a latent representation of the input sequence $\boldsymbol{x}$ [34, 26]. This computation requires previous training of the feature extraction using a large amount of binaural signals. It should however be emphasised that these signals do not need to be labeled and no assumption about the downstream task of SI prediction is made during feature computation. The SI is finally estimated by using the sequence $\boldsymbol{c}$ as input to a trained predicting function.

2.3. Predicting function

Given a new dataset of latent features $\boldsymbol{c}$ extracted from the trained VQ-CPC and associated intelligibility scores, we train a predicting function implemented as a lightweight neural network that controls global pooling [35], and a second neural network that makes a final prediction based off the pooled representation. This approach follows the "Pool" approach outlined in our previous work [25] inspired by sequence pooling strategies in low-data training of vision transformers [35].

For each frame in $\boldsymbol{c}$, a shared linear layer computes a scalar. All weightings are then collected and softmax is applied, forming normalised weightings. A weighted average of all frames is then computed, forming a global representation. This representation is fed into a multi-layer perceptron (MLP) and predicts the final intelligibility score, scaled to be between 0 and 1 [25].

The network is trained to minimise the mean-squared error (MSE) loss between the estimated and true speech intelligibility score. Building on our prior work, we tried more sophisticated predicting functions which incorporated deep convolutional networks and transformer architectures, but found the limited dataset size meant these more powerful architectures were prone to overfitting the training split. Hence, we found the simple predicting functions introduced in our earlier work to work best.

Table 1: *Overview of the Train and Test Datasets for both tracks of CPC1*

	Track 1		Track 2	
	Train	Test	Train	Test
Number of signals	4863	2421	3580	632
Total duration in hours	8.2	4.1	6.0	1.1
Number of algorithms	27	27	22	27
Number of listeners	10	10	9	10

3. Experiments

Training and evaluation of the proposed non-intrusive SI measure are done using the CPC1 dataset. The data consist of binaural signals that have been generated by convolving clean anechoic speech with various binaural room impulse responses (RIRs), adding noise at various signal-to-noise ratios (SNRs) and processing the resulting noisy and reverberant signal with speech enhancement algorithms designed for hearing aids. All signals have been labeled in terms of speech intelligibility in a

Table 2: *RMSE obtained when using the 3 considered measures*

	Track 1	Track 2
MBSTOI	28.51	26.61
$\overline{\text{SRMR}}_{\text{norm}}$	35.01	35.09
Proposed	38.93	39.41

listening test for which the audiogram of each listener has been measured. An overview of the dataset is presented in Table 1 but the interested reader can refer [28] for further details.

The proposed SI measure is evaluated on both track 1 and track 2 in order to examine the difference ein performance when applied to unknown algorithms or listeners. In track 1, all listeners and algorithms are represented in both training and test sets. In track 2, 5 of the listeners and 2 of the algorithms present in the test set are absent in from the training set. For all signals in the test set of track 2 algorithm, or listener, or both, are not present in the training set.

For both track 1 and track 2, we used 5-fold cross validation on the training set only to determine suitable settings for the proposed measure. In this case, the used folding is identical to the one used in the baseline software provided for CPC1 [33]. Settings yielding the best performance were then used on the complete test set, for both tracks. We report the results using cross-validation in the report and full training is used in the submitted scores. **In both tracks and for all test signals, the SI is predicted using only the target signal and the listener's audiogram. No data other than the one provided in the CPC1 dataset were used for training.**

4. Results

The performance of the proposed measure is assessed in terms RMSE (used to rank CPC1 submissions) between the measured and predicted SI. This performance is here benchmarked against the use of two other metrics. The considered metrics are the modified binaural STOI (MBSTOI) [36] and $\overline{\text{SRMR}}_{\text{norm}}$. $\overline{\text{SRMR}}_{\text{norm}}$ simply refers to the (non-intrusive) $\text{SRMR}_{\text{norm}}$ average over both input channels. Both use the signal obtained after applying the hearing loss model. For each of these 2 measures, the RMSE is computed after applying a sigmoidal mapping whose parameters were learned using the same training setup as for our proposed SI measure. In the case of MBSTOI, this is equivalent to using the baseline provided for CPC1. The obtained RMSE is presented in Table 2. **The final submission was done for both the proposed measure (E023) and the combination of hearing loss model, $\overline{\text{SRMR}}_{\text{norm}}$ and sigmoid mapping (E035)**.

5. Conclusion

The proposed measure does not outperform the baseline based on MBSTOI. As the proposed measure is non-intrusive, this was expected. However, it performs poorly compared to $\overline{\text{SRMR}}_{\text{norm}}$, despite using a more complex predicting function. This is disappointing considering the encouraging results obtained using VQ-CPC features. This might be due to the relatively small training dataset, compared to previous work. Further analysis, e.g., using results on the unseen test set, would be helpful to investigate this behaviour.

6. References

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

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

[3] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[4] N. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Nov. 1947.

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

[6] ANSI, "Methods for the calculation of the speech intelligibility index," American National Standards Institute, ANSI Standard S3.5–1997 (R2007), 1997.

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

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

[9] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 5085–5089.

[10] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Non-intrusive codebook-based intelligibility prediction," *Speech Communication*, vol. 101, pp. 85–93, 2018.

[11] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

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

[13] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, 1963.

[14] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension, and evaluation of a binaural speech intelligibility model," vol. 127, no. 4, pp. 2479–2497, 2010.

[15] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," vol. 127, no. 1, pp. 387–399, 2010.

[16] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and non-linearly processed binaural speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.

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

[18] R. Schädler, M. D. Hülsmeier, A. Warzybok, and B. Kollmeier, "Individual aided speech-recognition performance and predictions of benefit for listeners with impaired hearing employing FADE," *Trends in Hearing*, vol. 24, 2020.

[19] J. Roßbach, S. Röttges, C. F. Hauth, T. Brand, and B. T. Meyer, "Non-intrusive binaural prediction of speech intelligibility based on phoneme classification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, Jun. 2021, pp. 396–400.

[20] C. F. Hauth, S. C. Berning, B. Kollmeier, and T. Brand, “Modeling binaural unmasking of speech using a blind binaural processing stage,” *Trends in Hearing*, vol. 24, Jan. 2020.

[21] H. Hermansky, E. Variani, and V. Peddinti, “Mean temporal distance: Predicting ASR error from temporal properties of speech signal,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 7423–7426.

[22] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, “STOI-Net: A deep learning based non-intrusive speech intelligibility assessment model,” 2020, arXiv:2011.04292.

[23] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, “A data-driven non-intrusive measure of speech quality and intelligibility,” *Speech Communication*, vol. 80, pp. 84–94, 2016.

[24] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, “Nonintrusive speech intelligibility prediction using convolutional neural networks,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 1908–1920, Jul. 2018.

[25] A. F. McKinney and B. Cauchi, “Non-intrusive binaural speech intelligibility prediction from discrete latent representations,” *IEEE Signal Process. Lett.*, 2022, to Appear.

[26] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2019, arXiv:1807.03748.

[27] B. van Niekerk, L. Nortje, and H. Kamper, “Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge,” 2020, arXiv:2005.09409.

[28] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros-Muñoz, “Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing,” in *Proc. Interspeech*, Brno, Czech Republic, Aug. 2021.

[29] T. Baer and B. C. Moore, “Effects of spectral smearing on the intelligibility of sentences in noise,” *J. Acoust. Soc. Am.*, vol. 94, no. 3, pp. 1229–1241, 1993.

[30] ——, “Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech,” *J. Acoust. Soc. Am.*, vol. 95, no. 4, pp. 2277–2280, 1994.

[31] B. C. Moore and B. R. Glasberg, “Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech,” *J. Acoust. Soc. Am.*, vol. 94, no. 4, pp. 2050–2062, 1993.

[32] M. A. Stone and B. C. Moore, “Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses,” *Ear and Hearing*, vol. 20, no. 3, pp. 182–192, 1999.

[33] J. Barker, S. Graetzer, and T. Cox, “Software to support the 1st clarity enhancement challenge [software and data collection],” 2021, https://doi.org/10.5281/zenodo.4593856.

[34] V. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” 2018, arXiv:1711.00937.

[35] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, “Escaping the big data paradigm with compact transformers,” 2021, arXiv:2104.05704.

[36] 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.