Non-intrusive speech intelligibility prediction from binaural signals processed for hearing aid users

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 signal and audiogram of the listener as input. The proposed measure relies on three steps, namely a hearing-loss model, a VQ-CPC 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 training and testing sets defined for both tracks of the first Clarity Prediction Challenge (CPC1). Results show that despite encouraging results obtained by training VQ-CPC on large amounts of data, the measure is here instead outperformed by the considered benchmark.

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 ($\text{SRMR}_{\text{norm}}$) [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

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.

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 opportunity, the work presented in this paper 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 measure is evaluated in terms of root mean-squared error (RMSE) and correlation and benchmarked against intrusive and non-intrusive measures that are combined with the same hearing loss model and a simple (trained) sygmoidal mapping. Results show that the proposed measure, despite previous encouraging results of VQ-CPC features, is outperformed by the considered benchmark, including the CP1 baseline, in terms of both correlation and RMSE.

The remainder of this paper 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 and their discussion are presented in Section 4. Section 5 concludes the paper.

2. Proposed approach

The non-intrusive SI measure that is proposed in this paper aims to evaluate speech intelligibility. Specifically, the SI of a listener with hearing loss when exposed to a noisy reverberant signal processed through some 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, presented in the following subsections.

2.1. Hearing loss model

SI is largely dependent on the type and severity of the hearing loss in the target listener. 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 VQ-CPC 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 $2 \cdot W$:

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

resulting in the time-ordered sequence of T vectors:

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

The feature computation results in the sequence:

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

where $\boldsymbol{c}_t$ denotes the vector of length K feature coefficients extracted from the t^{th} frame. The feature extractor 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 training of a 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 or extractor training. The SI is finally estimated by using the sequence $\boldsymbol{c}$ as input to a trained predicting function.

We use a larger VQ-CPC model than in our original work [25]. Like before, we train on windows of audio consisting of $T = 40960$ samples, and with an encoder model comprised of convolutional blocks formed of a one dimensional convolutional layer with 256 filters, a dropout layer [35], batch normalisation [36], and a rectified linear unit (ReLU) activation function. However, the number of blocks is increased from 5 to 7, with strides $[5, 4, 2, 2, 2, 1, 1]$ and kernel sizes of $[10, 8, 4, 4, 4, 1, 1]$.

The VQ codebook contains 512 codewords each of dimensionality 128, using a k-means strategy to select codewords given the encoder output [34]. The aggregator network is implemented as a two-layer gated recurrent neural network (GRU) with 128 hidden channels – identical to the VQ codebook dimensionality. The InfoNCE [26] loss is computed using 10 negative samples and $k = 12$ steps. We also apply limited data augmentation, including random channel and polarity swapping, additive noise and random audio gain [25]. Despite these augmentations affecting the intelligibility of the signal, this does not affect VQ-CPC training as it uses $\boldsymbol{x}$ alone to train – not any intelligibility labels.

These parameters gives an effective frame length and hop size of the whole model of 25 ms and 10 ms respectively, giving a sequence $\boldsymbol{c}$ with $K = 128$. Overall, the parameters of the VQ-CPC model are identical to those in our original work [25] with the exception of an expanded encoder network.

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 [37], and a second neural network that makes a final prediction based off the pooled representation. This approach follows the "Pool" approach outlined

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

in our previous work [25] inspired by sequence pooling strategies in low-data training regimes of vision transformers [37]. Our previous work also proposed a variation that predicted intelligibility per frame, then averaged the per-frame scores to produce a final prediction. This had an advantage of allowing for a per-frame breakdown of the perceived intelligibility, however it ultimately had worse performance than the pooling strategy [25]. Hence, as we are aiming at high performance in a competition scenario, we selected the pooling approach.

For each frame in $\boldsymbol{c}$, a shared linear layer computes a scalar value. All weightings are then collected and softmax function is applied, forming normalised weightings. A weighted average of all frames is then computed, subsequently creating 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] using the sigmoid activation function. The weighting mechanism allows the predicting function to assign relative importance in predicting speech intelligibility to each frame, rather than simply taking a naïve average of all frames.

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 dataset due to its comparatively small size. Hence, we found the simple predicting functions introduced in our earlier work to work best. We consider two variations, one with a hidden layer size of 256 and a second with a hidden layer size of 512, which are denoted "Small" and "Large" predictors respectively.

3. Experiments

3.1. Used datasets

Training and evaluation of the proposed non-intrusive SI measure is 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 listening test and in which the audiogram of each listener has also 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 in 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, five of the listeners and two 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 the algorithm, or listener, or both, are not present in the training set. In our previous work, we found the VQ-CPC feature extractor to transfer well to unseen noise types [25] and unseen speakers (following evaluation on the LibriSpeech corpus [38]) so we hypothesized that this advantage would translate well to unseen listener audiograms and algorithms.

For both track 1 and track 2, and for all test signals, the SI is predicted using only the target signal and the listener's audiogram. No data, other than those provided in the CPC1 dataset, was used for training.

3.2. Benchmark and figures of merit

The performance of the proposed measure is assessed in terms of RMSE (used to rank CPC1 submissions) and of Pearson's correlation coefficient (LCC) between the measured and predicted SI. This performance is benchmarked against the use of existing signal-based metrics computed from $x_m(n)$, i.e., the output of the hearing loss simulation.

Four signal-based measures are used as benchmark in this paper, namely the modified binaural STOI (MBSTOI) [39], STOI [7], its non-intrusive extension NISTOI [9], and $\text{SRMR}_{\text{norm}}$ [12]. All four measures are computed from the output of the hearing loss model and after, with the exception of MBSTOI, processing $x_m(n)$ into a single-channel signal using a blind binaural preprocessing stage (BSIM20) [20]. For all four signal-based measures, speech intelligibility is computed after applying a sigmoidal mapping whose parameters are learned using the CPC1 training set. In the case of MBSTOI, this is equivalent to using the baseline provided for CPC1.

4. Results

The performance of the considered measures are depicted in Fig. 1. Unsurprisingly, the intrusive measure (MBSTOI and STOI) perform best in terms of both RMSE and LCC. The quite small degradation in performance observed when comparing STOI to MBSTOI, seems to suggest the the BSIM20 model, here applied to reduce the binaural signal to a single-channel signal while preserving binaural information, could be a suitable approach in future designs on intelligibility measures. One should however consider that even the performance of MBSTOI on this challenging task is far from perfect, with its best performance obtained on Track 1 with an RMSE of 28.52 and an LCC of 0.62.

Though the lower performance of non-intrusive measures was expected, the considered measures yield very low correlation, with a maximum of 0.25 for both NISTOI and $\text{SRMR}_{\text{norm}}$. The combination of VQ-CPC features and with both the Large and Small predictor completely failed. Several hypothesis can be made to explain this failure. Concerning the poor performance of all non-intrusive measures, this is possibly due to their inadequacy in capturing the particularity of the hearing loss in each listener. One potential solution is to condition the VQ-CPC feature extractor directly on the listener's audiogram as an auxiliary input to $\boldsymbol{x}$, potentially improving the quality of the resulting latent representations. A similar approach could also be taken with the predicting function, again taking the audiogram as an auxiliary input alongside $\boldsymbol{c}$.

More importantly in the context of this paper is the failure of the VQ-CPC based measures denoted as Small and Large.

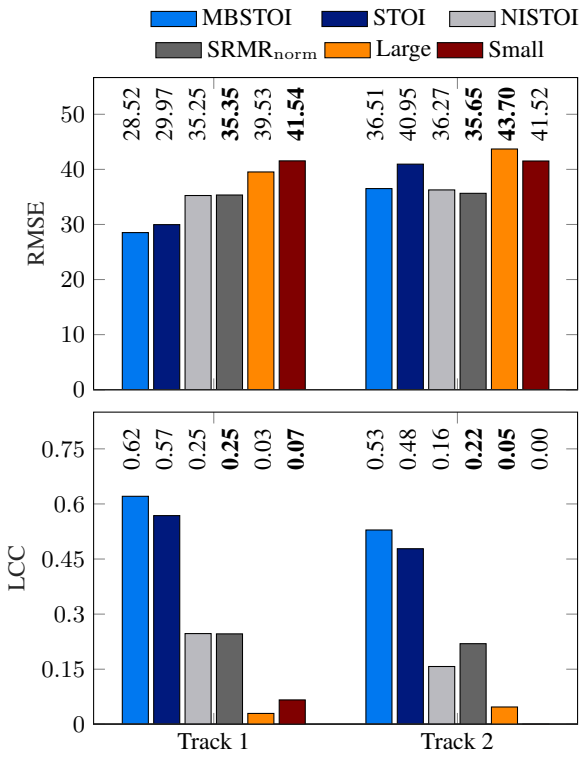

Figure 1: *Performance of the proposed non-intrusive measure using VQ-CPC features and either the Small or Large predictor, and considered benchmark signal-based measures. Blue indicates intrusive measures (MBSTOI and STOI) and grey indicates non-intrusive measures (NISTOI and SRMR$_{\text{norm}}$). Bold typeface indicates the measures whose output were submitted to CPC1.*

The most likely explanation for this failure is the small size of the training dataset. Indeed, previous work using a similar measure used several hundreds of hours of training data [25]. This was however done at the cost of using an intrusive measure to label the signals rather than, as in CPC1, real measurements of intelligibility. Future works, possibly in future prediction challenges, might allow us to use both large datasets and real intelligibility labels, though we acknowledge that such an affair is most likely to be costly to undertake.

A more cost-efficient solution is to train the VQ-CPC on large amounts of binaural audio which does not have to have intelligibility scores associated with it. However, the limited amount of labelled data means that the predicting functions would still be forced to use a significantly smaller dataset. This limits the selection of architectures that can be used in the predicting function. This is because potentially more powerful architectures are prone to overfitting in data-limited scenarios, making their use untenable. Another potential solution is to use more aggressive data augmentation, expanding the effective size of the dataset.

5. Conclusion

This paper proposed to apply the combination of a hearing loss model, a VQ-CPC feature extractor and trained predictor to predict the intelligibility from binaural signals processed using hearing aids algorithms. This prediction is done non-intrusively, using only the target signal and the audiogram of the target listener as input. The evaluation is conducted on the CPC1 dataset and this measure was submitted to this challenge. The proposed measure does not outperform the MBSTOI-based baseline and, perhaps more importantly, fails even when compared to the other considered non-intrusive benchmark measures. 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 in our prior work. This might be due to the relatively small training dataset compared to the dataset used in our previous work. Future research, including in future prediction challenges, would be helpful to investigate this behaviour further.

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] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, arXiv:1502.03167.

[37] 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.

[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 5206–5210.

[39] 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, Sep. 2018.