Non-Intrusive Binaural Speech Intelligibility Prediction From Discrete Latent Representations

Alex F. McKinney Student Member, IEEE and Benjamin Cauchi Member, IEEE

Abstract—Non-intrusive speech intelligibility (SI) prediction from binaural signals is useful in many applications. However, most existing signal-based measures are designed to be applied to single-channel signals. Measures specifically designed to take into account the binaural properties of the signal are often intrusive – characterised by requiring access to a clean speech signal – and typically rely on combining both channels into a single-channel signal before making predictions. This paper proposes a non-intrusive SI measure that computes features from a binaural input signal using a combination of vector quantization (VQ) and contrastive predictive coding (CPC) methods. VQ-CPC feature extraction does not rely on any model of the auditory system and is instead trained to maximise the mutual information between the input signal and output features. The computed VQ-CPC features are input to a predicting function parameterized by a neural network. Two predicting functions are considered in this paper. Both feature extractor and predicting functions are trained on simulated binaural signals with isotropic noise. They are tested on simulated signals with isotropic and real noise. For all signals, the ground truth scores are the (intrusive) deterministic binaural STOI. Results are presented in terms of correlations and MSE and demonstrate that VQ-CPC features are able to capture information relevant to modelling SI and outperform all the considered benchmarks – even when evaluating on data comprising of different noise field types.

Index Terms—Non-intrusive speech intelligibility prediction; self-supervised representation learning; contrastive predictive coding.

I. INTRODUCTION

Speech intelligibility (SI) prediction aims to predict the ability of an average listener to comprehend speech within a signal – potentially corrupted by noise, reverberation or processing artefacts. SI is defined as the number of words or phonemes that can be correctly identified by assessors. It is often reported using the speech reception threshold (SRT) as a latent representation of the input binaural signal used as input to the STOI measure. BSTOI was later refined into the deterministic BSTOI (DBSTOI) that produces a deterministic output. Both BSTOI and DBSTOI are intrusive. A non-intrusive measure has been proposed in [17], where the blind binaural preprocessing stage from [18] is used to process the binaural signal into a single-channel signal that is then input to an automatic speech recognizer (ASR). The SI is finally predicted by applying a trained mapping between the mean temporal distance (MTD) – a representation of the ASR error [19] – and the SRT.

Some measures do not rely on any model of the auditory system and input features to a predicting function that needs to be trained. Such methods include the use of both short- and long-term features input to a classification and regression tree [20] or the use of STOI like features as input to a convolutional neural network [21]. The measure proposed in this paper applies a similar approach but uses features that are computed as a latent representation of the input binaural signal using a combination of contrastive predictive coding (CPC) [22].
The resulting VQ-CPC features are input to a predicting function. Two predicting functions are considered in this paper to highlight the capacity of the proposed features to capture useful and accessible information and to show that our measure is competitive.

The remainder of this paper is structured as follows. The experimental setup, including the used datasets of simulated data and considered benchmark, are described in Section III. The results are presented in Section IV and Section V concludes the paper.

II. PROPOSED METHOD

The proposed method aims at estimating the speech intelligibility from an M channel signal \(x_m(n)\), of length \(N\) and sampling frequency \(f_s\), where \(m\) and \(n\) denote the channel index and sample index respectively. Such a signal can be modelled as:

\[
x_m(n) = s(n) * h_m(n) + v_m(n),
\]

where \(s(n)\) denotes the anechoic speech signal, \(h_m(n)\) denotes the room impulse response (RIR) between the speech source and the microphone and \(v_m(n)\) denotes an additive noise signal. Aiming at non-intrusive prediction, the proposed method estimates the speech intelligibility from \(x_m(n)\) only without knowledge of \(s(n)\). This prediction relies on first computing a sequence of features to be input to the predicting function.

A. Feature computation

The microphone signal is divided into \(T = [N/H]\) 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 \(M \cdot W\):

\[
x_t = [x_0(tH), x_0(tH + 1), \ldots, x_0(tH + W - 1),
\]

\[
\ldots
\]

\[
x_{M-1}(tH), x_{M-1}(tH + 1), \ldots, x_{M-1}(tH + W - 1)]^T,
\]

resulting in the time-ordered sequence of \(T\) vectors:

\[
x = \{x_0, x_1, \ldots, x_{T-1}\}. \tag{3}
\]

The feature computation results in the sequence:

\[
c = \{c_0, c_1, \ldots, c_{T-1}\}, \tag{4}
\]

where each vector of length \(K\) is defined as:

\[
c_t = [c_t(0), c_t(1), \ldots, c_t(K-1)]^T, \tag{5}
\]

where \(c_t(k)\) denotes the \(k\)th feature coefficient extracted from the \(t\)th frame. The feature extraction is typically designed such that \(K < M \cdot W\) and learns to extract sequences \(c\) that maximise the mutual information between the input and output sequences:

\[
I(x; c) = \sum_{x,c} p(x,c) \log \left( \frac{p(x|c)}{p(x)} \right). \tag{6}
\]

To do so, VQ and CPC methods are used to compute the sequence \(c\) as a latent representation of the input sequence \(x\). The computation of these VQ-CPC features consists of three main components: a non-linear encoder, a VQ codebook, and an autoregressive aggregator.

First, the non-linear encoder \(f(\cdot)\) maps \(x\) to an intermediate latent representation \(\tilde{z}\):

\[
f(x) = \tilde{z} = \{\tilde{z}_0, \tilde{z}_1, \ldots, \tilde{z}_{T-1}\}, \tag{7}
\]

where \(\tilde{z}_t\) denotes the \(t\)th vector, each of length \(E\). VQ is applied to map each vector in \(\tilde{z}\) to an embedding vector from a finite codebook \(C\) yielding the sequence:

\[
z = \{z_0, z_1, \ldots, z_{T-1}\}, \tag{8}
\]

where each \(t\)th vector \(z_t\) is computed as:

\[
z_t = q(\tilde{z}_t) = \arg \min_{e_i \in C} ||\tilde{z}_t - e_i||_2 \tag{9}
\]

Where \(e_i\) denotes the \(i\)th in the \(C\) embedding vectors of the codebook. Finally, an autoregressive aggregator \(g(\cdot)\) is applied to compute each vector from the sequence in \((4)\) as:

\[
c_t = g(z_{t\leq t}). \tag{10}
\]

B. VQ-CPC training

Training of \(f(\cdot)\), \(q(\cdot)\) and \(g(\cdot)\) is conducted end-to-end to maximize the mutual information defined in \((6)\). The proposed approach follows the method in \((22)\) with additional loss terms to support the added VQ codebook \((20)\). To encourage shared information to be encoded, each vector \(c_t\) is used to predict \(z_{t+k}\) for up to \(k\) steps in the future. However, rather than modelling the distribution \(p(x_{t+k}|c_t)\), the proposed method models the density ratio defined as:

\[
\sigma_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}. \tag{11}
\]

The density ratio \(\sigma_k(x_{t+k}, c_t)\) may be unnormalized and, in this paper, is computed as:

\[
\sigma_k(x_{t+k}, c_t) = \exp \left( z_{t+k}^T W_k c_t \right), \tag{12}
\]

where \(W_k\) denotes a learned linear projection and \(z_{t+k}\) is the output of the encoder corresponding to \(x_{t+k}\), used as a proxy for more efficient computation of the ratio. Using
this definition, the encoder and aggregator are trained by
minimising the InfoNCE loss $\mathcal{L}$, based on noise-contrastive
estimation and importance sampling:

$$
\mathcal{L} = \beta \cdot \mathcal{L}_{\text{vq}} + \frac{1}{k} \sum_{l=1}^{k} \mathcal{L}_l,
$$

(13)

where $\mathcal{L}_{\text{vq}}$ denotes the weighted VQ commitment loss defined as:

$$
\mathcal{L}_{\text{vq}} = \frac{1}{T} \sum_{t=0}^{T-1} \left( \| \tilde{z}_t - \text{sg}[e_t] \|_2^2 \right)
$$

(14)

where $e_t$ is the corresponding embedding vector of $\tilde{z}_t$ and
$\text{sg}[-]$ is the stop-gradient operator [26] and:

$$
\mathcal{L}_k = - \mathbb{E}_X \left[ \log \frac{\sigma_k(x_{t+k}; c_t)}{\sum_{x_j \in X} \sigma_k(x_j; c_t)} \right],
$$

(15)

where $X$ is a set of many negative samples drawn from
$p(x_{t+k})$ and one positive sample drawn from $p(x_{t+k}; c_t)$
[22]. The codebook embedding vectors are updated using
exponential moving averages (EMA) as described in [26].

C. Intelligibility score predictor

The computation of the VQ-CPC features does not rely on
any assumptions about the downstream task for which these
features are used. In this paper, the sequence $c$ is input to a
predicting function for the purpose of SI prediction. Two
different predicting functions are considered.

The first considered predicting function uses each vector $c_t$
as input to a single shared linear layer in order to compute a
per-frame score. The score assigned to the complete sequence
is the mean of the scores computed from each vector. This
simple predicting function is used to demonstrate how easily
accessible information about SI is when using the VQ-CPC
features. This predicting function is referred to as “Small” in
the remainder of the paper.

The second considered predicting function first builds a
global representation using sequence pooling (SeqPool) meth-
ods originally used for the classification of images [27]. A
global representation is computed by applying SeqPool to each
vector in the sequence $c$. In this case SeqPool inputs each
vector $c_t$ to a linear layer that outputs a scalar before applying
softmax to the computed scalars, forming weightings for each
frame. The weighted sum of each vector is then computed,
forming the global representation. This global representation
is then input to a small multi-layer perceptron (MLP) to
compute the estimated speech intelligibility score assigned
to the sequence $c$. This predicting function is referred to as
“Pool” in the remainder of the paper.

Both Small and Pool are trained to minimise the mean-
squared error (MSE) between the estimated and true speech
intelligibility scores (see Section III).

III. EXPERIMENTAL SETUP

A. Generated datasets

For training of both the VQ-CPC model and the predict-
ing functions, training and development datasets of binaural
signals are generated. All signals have a sampling frequency
$f_s = 16$ kHz and are generated as per [1]. The clean anechoic
speech is extracted from either the 360 hour training set or
the 5 hour development set from the LibriSpeech corpus [28].
Reverberant speech is generated by convolving each utterance of clean speech with a binaural RIR (BRIR) randomly selected
from the Aachen Impulse Response Database [29]. For each
reverberant utterance, two different noise segments of the same
length are selected from the noise signals in the MUSAN
database [30]. These two signals are used to generate the two-
channel noise signal of a spherically isotropic noise field using
the method from [31]. Finally, this generated noise signal is
added to the reverberant signal after being scaled to a chosen
signal-to-noise ratio (SNR), randomly selected between -10 dB
and 30 dB, measured in the first channel according to [32].
In the training and development sets, this process is repeated
three times for each clean speech utterance.

Additionally, two testing datasets are generated, hereafter
denoted “Testiso” and “Testreal”. The signals in Testiso are
generated using the same method, as well as noise and
BRIR datasets, as for the training and development sets but
using clean speech from the test split of the LibriSpeech
corpus. The signals in Testreal are generated by convolving
the speech signals used as target utterances in the first Clarity
Challenge [33, 34] with BRIRs randomly selected from the
BRIRs available in [35] recorded in either a cafeteria or a
courtyard. In this case, two-channel noise signals recorded at
the same location are used and added to the reverberant signals
with an SNR randomly selected and measured.

A total of 1090.8, 16.2, 5.4 and 10.4 hours of data are
generated in the training, development, Testiso and Testreal
dataset, respectively. Labelling this large amount of data in
terms of intelligibility would be a daunting task and the
experiments aim mostly at evaluating the use of the proposed
features. Consequently, we labeled all signals with an intrusive
measure known to highly correlate with intelligibility and the
ground truth is here defined as the DBSTOI computed using
the clean reverberant signal and the noisy reverberant signal
as input [16].

B. Parameters of proposed method

For training, we use $x$ of length $T = 40960$ as input to
the encoder $f(\cdot)$. The encoder has a frame length and hop
size of 25 ms and 10 ms respectively. It is implemented as a
series of five convolutional blocks, each consisting of a one-
dimensional convolutional layer with 256 filters, a dropout
layer [36], batch normalisation [37] and the rectified linear
unit (ReLU) activation function. The strides for each block
are $[5, 4, 2, 2, 2]$ and the kernel sizes are $[10, 8, 4, 4, 4]$. VQ is
applied using a codebook of 512 vectors of dimensionality
128, with the commitment loss defined as in [14]. The aggre-
gator $g(\cdot)$ is implemented as a two-layer gated recurrent neural
network (GRU) [38] with 128 hidden channels. Hence, in our
experiments, $K = E$. The InfoNCE loss is computed using 10
negative samples and $k = 12$ steps. Augmentation is applied
as random channel and polarity swapping, additive noise and
random audio gain. All resulting sequences $c$ (with $K = 128$)
The hidden layer uses ReLU as its non-linearity and the output of trainable network weights in the VQ-CPC model is 

functions were implemented in PyTorch [39]. The total number of intelligibility prediction. To this end, their performance is 

VQ-CPC features to represent information useful for speech 

functions. The experiments aim to quantify the ability of 

between the ground truth and the output of the predicting 

terms of Pearson’s correlation coefficient (LCC) and MSE 

C. Benchmark and figures of merit

The performance of the proposed VQ-CPC is measured in 

terms of Pearson’s correlation coefficient (LCC) and MSE 

between the ground truth and the output of the predicting 

functions. The experiments aim to quantify the ability of 

VQ-CPC features to represent information useful for speech 

intelligibility prediction. To this end, their performance is 

compared with the use of mel-spectrogram (Mel), with deltas 

and double-deltas, and with envelopes extracted in third-

octave bands (TOB) similarly as used in [21]. All features 

are extracted from either the first channel (mono), concatenated 

from both channels (binaural) or extracted from the single-

channel signal computed using a blind binaural preprocessing 

stage [18] (BSIM20). The type of signal is indicated in 

subscript in the following. All features are computed using 

the complete VQ-CPC model trained on the 

training set. 

The Small predicting function consists of a single layer 

mapping each feature vector of length $K$ to a single element 

followed by the sigmoid activation function. The Pool predict-

or is implemented as a single shared linear layer to compute 

the weighting and a MLP with one hidden layer of size $2K$. 

The hidden layer uses ReLU as its non-linearity and the output 

layer consists of a single element followed by the sigmoid 

activation function. 

Training and testing of the VQ-CPC model and predicting 

functions were implemented in PyTorch [39]. The total num-

ber of trainable network weights in the VQ-CPC model is 

$1.74 \times 10^6$. 

IV. RESULTS

All results are depicted in Fig. 1. On Testiso, VQ-CPC 

features yield the best performance regardless of the type of 

signals from which they are computed, when using either the 

Small or the Pool predicting function. Using Small and VQ-

CPC features yields a LCC 0.84 and an MSE of -20.7 dB. Using Pool and VQ-CPC features yields a LCC of 0.94 and 

an MSE of -22.7 dB. In contrast with the other considered 

features, the difference in performance between the use of 

Small and Pool is rather modest. Success applying Small 
suggests that the VQ-CPC features contain easily accessible 

information about the intelligibility of speech. 

On Testreal, the performance of all combinations of features 

and predicting functions decreases, as expected. It can however 

be noted that the TOB features, that performed the least 
satisfactorily on the less challenging Testiso, outperform Mel 
on Testreal. This seems to confirm their suitability in realistic scenarios [21]. The proposed VQ-CPC features remain the best 

performing of the considered features. Using VQ-CPC features 

computed from binaural signals as input to the Pool predicting 

function yields LCC of 0.81 and an MSE of -20.2 dB. 

A more powerful predictor such as STOI-Net [12] could be 
improve performance further, but we emphasize the purpose 
of our study is to show that good performance can be obtained 

with VQ-CPC features alone. Though the VQ-CPC were 

here proposed to predict intelligibility from binaural signals, 

the difference in performance between VQ-CPCmono, VQ-

CPCBSIM20 and VQ-CPCbinaural is modest. Further exper-

imentation, e.g., using intelligibility scores as ground truth 

rather than an intrusive measures, are needed to determine if 

the VQ-CPC features do capture information such as binaural 
cues. Regardless, the difference in network size between the 

various VQ-CPC models is negligible. 

V. CONCLUSION

This paper proposes to use VQ-CPC features as input to a 

trained neural network to non-intrusively predict intelligibil-

ity from binaural signals. The performance of the proposed 

measure is assessed in terms of correlation and MSE. Results 

show that the VQ-CPC features are effective in encoding 

readily accessible information relevant for SI prediction and 

the features outperform all considered benchmarks. This is 

despite VQ-CPC features not relying on any assumptions 

about the downstream task of SI prediction.
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] 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.

[3] 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.

[4] 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.

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

[6] C. H. Taal, R. C. Hendrikx, 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, Sept. 2011.

[7] J. Taglia 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.

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

[9] 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.

[10] 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.

[11] R. Schädel, 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. 20, 2020.

[12] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, “STOI-2vec: Unsupervised pre-training for speech recognition,” 2019, arXiv:1904.05862.

[13] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised deep learning of discrete speech representations,” 2020, arXiv:1910.05453.

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

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

[16] 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.

[17] M. Jeub, M. Schaefler, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in Proc. IEEE Intl. Conf. Digital Signal Processing (DSP), Santorini, Greece, July 2009.

[18] D. Snyder, G. Chen, and D. Povey, “MUSAN; a music, speech, and noise corpus,” 2015, arXiv:1510.08484v1.

[19] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” J. Acoust. Soc. Am., vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[20] ITU-T, “Objective measurement of active speech level.” Mar. 1993.

[21] S. Graeter, 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.

[22] S. Graeter, M. 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, Apr. 2022.

[23] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,” EURASIP Journal on Advances in Signal Processing, vol. 2009, pp. 1–10, July 2009.

[24] 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.

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

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, arXiv:1412.3555.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, 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,” 2019, arXiv:1912.01703.