# Voice over LTE Quality Evaluation Using Convolutional Neural Networks

Thomas Gorman
*School of Computing, Engineering and Built Environment*
*Glasgow Caledonian University*
Glasgow, United Kingdom
tgorma200@caledonian.ac.uk

Hadi Larijani
*School of Computing, Engineering and Built Environment*
*Glasgow Caledonian University*
Glasgow, United Kingdom
H.Larijani@gcu.ac.uk

Ayyaz-Ul-Haq Qureshi
*School of Computing, Engineering and Built Environment*
*Glasgow Caledonian University*
Glasgow, United Kingdom
Ayyaz.Qureshi@gcu.ac.uk

***Abstract*—Modern packet-switched networks are increasingly capable of offering high-quality voice services such as Voice over LTE (VoLTE) which have the potential to surpass the Public Switched Telephone Network (PSTN) in terms of quality. To ensure this development is sustained, it is important that suitable quality evaluation methods exist in order to help measure and identify the effect of network impairments on voice quality. In this paper, a single-ended, objective voice quality evaluation model is proposed, utilizing a Convolutional Neural Network with regression-style output (CQCNN) to predict mean opinion scores (MOS) of speech samples impaired by a VoLTE network emulation. The results of this experiment suggest that a deep-learning approach using CNNs is highly successful at predicting MOS values for both narrowband (NB) and super-wideband (SWB) samples with an accuracy of 91.91% and 82.50% respectively.**

***Index Terms*—Voice quality, VoLTE, CNN, MOS, SWB, NB, Deep Learning**

## I. Introduction

Voice over IP (VoIP) has grown rapidly since its inception, leading many to believe it will eventually succeed the Public Switched Telephone Network (PSTN) as the preferred voice communication technology [1]. VoIP's potential as a cost-effective, high quality alternative to the circuit-switched PSTN has been capitalised on by mobile telecommunications networks.

Voice over LTE (VoLTE) is an implementation of VoIP which utilizes the high performance of 4G mobile network architectures to offer packet-switched calling at qualities equal to or exceeding those of the PSTN [2]. This is achieved using real-time optimized technologies such as the IP Multimedia Subsystem (IMS) [3] and modern adaptive codecs which can maximize call quality and reliability even in challenging network conditions [4].

VoLTE has used adaptive multi-rate codecs since the service launched. While the AMR-WB (G722.2) codec was used initially, the recently standardized Enhanced Voice Services (EVS) codec has been rolled out due to its improved performance and features such as channel-awareness [5] and forward error correction (FEC) [6] which can help audio streams to recover from networks with a high packet-loss rate (PLR). Though the IETF and WebRTC standardized adaptive multi-rate codec Opus [7], is not utilized directly in VoLTE services, it is used in modern over-the-top mobile VoIP applications such as "WhatsApp" and has similar FEC capabilities. These codecs both offer Narrowband (NB), Wideband (WB), Super-wideband (SWB) and Full-band (FB) mode-sets at both constant and variable bitrates. They are both capable of utilizing the full potential of modern mobile networks to deliver HD voice and clearly outperform competitors in industry tests [2].

Despite the potential evident in these technologies, there are considerable issues faced by VoLTE. Traditional PSTN networks rely on an end-to-end circuit over a predominantly wired physical infrastructure. This leads to a robust, reliable network with high availability of 99.999% or just 5 minutes downtime per year [8]. Packet-switched topologies, such as the 4G LTE networks on which VoLTE operates, are often connected via different physical media and are subsequently less reliable. These networks are also subject to impairment factors such as delay, jitter or packet loss which can have a significant impact on call and voice quality. A variety of evaluation methods exist to measure call quality in these environments and can be categorized as subjective or objective.

Subjective call quality evaluation involves experiments where human subjects assign the quality of the call or listening experience an absolute category rating (ACR) from 1 (poor quality with very annoying impairment) to 5 (excellent quality with imperceptible impairment). The mean of these ratings is then taken to give a mean opinion score (MOS), giving an extremely accurate measurement of quality [9]. Such experiments are hard to conduct in an unbiased nature and require a large and varied pool of test subjects.

Objective methods evaluate the quality of voice samples through the measurement of signal data across the network [10]. Such methods may be described as intrusive or non-intrusive depending on the extent of network access and signal data they required in their analysis. Intrusive methods such as ITU-T's Perceptual Objective Listening Quality Analysis tool

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

(POLQA) specified in recommendation P.863 [11] and its predecessor, recommendation P.862 Perceptual Evaluation of Speech Quality (PESQ) [12] are algorithms which compare a clean reference signal to degraded version after it has transited the network. While these methods produce accurate results, their algorithms are often implemented in expensive equipment and are unusable without live network traffic and pervasive network access. Beyond this, PESQ is only recommended for use with NB/WB signals – making it unsuitable for modern VoLTE networks. POLQA supports SWB signal analysis but is a proprietary technology, available almost exclusively in its costly hardware form.

Non-intrusive standards for call quality assessment can evaluate signals by detecting distortions, interruptions and unnatural sounding speech [13] without the need for a reference signal. Single-ended methods are generally less accurate than intrusive methods but can be effective in giving a general indication of voice quality when a reference is unavailable. ITU-T's recommendation P.563 is the only standardised, single-ended algorithm for call quality measurement and is rendered ineffective in a VoLTE environment as it is only designed to process 16-bit NB signals. Previous research [14, 15] has suggested that Artificial Intelligence (AI), particularly Artificial Neural Networks (ANNs) could produce reliable, single-ended tools for evaluating voice quality in a variety of network contexts. Artificial Neural Networks are a mathematical attempt to closely model the functions of biological neural networks. The two principal attributes that these models describe are the architecture and the functional properties, or neurodynamics, of such networks. Each neuron, or node, in an ANN is "excited" by an input to generate an output value. This activation potential($u$) is given by subtracting the bias/activation threshold $(\theta)$ from the sum total of the input values $(x_i)$ multiplied by their corresponding synaptic weights $(w_i)$ as described below:

$$u = \sum_{i=1}^{n} w_i \cdot x_i - \theta \tag{1}$$

The output value $(y)$is then calculated by applying the activation function $(g(x))$ to this activation potential $(u)$.

$$y = g(u) \tag{2}$$

When arranged into architectures with enough neurons and layers, these models can be used to approximate a large variety of functions and learn features from its input data in order to produce some desired output. ANNs use supervised machine learning methods that require a labelled dataset to achieve this task.

Deep Learning, driven by advances in computational technology such as more powerful GPUs and new techniques which solved long-standing issues of overfitting in deep neural architectures, has allowed for increased performance across all major areas in machine learning [16]. Convolutional Neural Networks (CNNs) have seen a resurgence in popularity since the advent of Deep Learning. CNNs work using sparsely connected layers of neurons which mimics the architecture of the human visual cortex. Matrix convolutions are performed by passing a 2D kernel over a 2D input to extract features. This is process is described by the equation:

$$S(i,j) = (I*K)(i,j) = \sum_{m}\sum_{n} I(m,n)K(i-m,j-n) \tag{3}$$

CNNs natively preserve the 2D form of its input, making it ideal for use with images. Convolutional models have also been successfully used for numerous audio applications such as text-to-speech [17], voice recognition [18] and audio enhancement [19] but they have not been utilised in the domain of call quality analysis. This research aims to assess the viability of a convolutional neural network (CNN) and deep learning to predict MOS values for samples typical of a VoLTE environment.

The primary contributions of this paper are:

- The development of a novel, single-ended, objective call quality evaluation tool (CQCNN) that outperforms current standards using a CNN architecture and deep learning.
- The extension of objective, single-ended call quality evaluation to SWB VoLTE signals.
- The modification of the VoxForge[20] dataset via realistic VoLTE network emulation to produce a large and representative dataset of paired samples for use in further objective call quality evaluation research.
- The performance of the proposed model (CQCNN) is critically investigated and compared to existing ITU-T objective standards including the state-of-the-art POLQA algorithm as well as similar, previously proposed methods of quality evaluation.
- Empirical analysis of the effects of network impairment factors and sample quality related to the proposed model's predictions.

The remainder of the paper is organized as follows:
Preliminary work to examine any related research is detailed in Section-II. The methodology employed in this paper to develop the proposed model is discussed in detail in Section-III. The results of this research are presented with accompanying analysis in Section-IV and conclusions derived from these are delineated in Section-V with further research directions also suggested.

## II. Related Work

### A. VoIP Call Quality Evaluation

A wide-range of techniques for evaluating call quality in packet-switched networks have been proposed in previous work. Parametric methods building on ITU-T's estimative standard (E-model) have been used to evaluate VoLTE call quality in [27]. [28] demonstrates a hybrid model which uses these non-intrusive measurements as an initial quality

estimate and refines them via intrusive measurements for a more accurate real-time quality monitoring system. [29] demonstrates the effectiveness of various Machine Learning techniques for the purpose of call quality evaluation with Ordinal Logistic Regression (OLR) achieving the best accuracy (61%) in predicting PESQ-generated MOS scores. Artificial Neural Networks were excluded from their analysis due to the lack of suitable datasets identified by the authors.

### B. VoIP and Deep Learning

Artificial Neural Networks are an extremely appropriate technology for assessing voice quality due to their ability to accurately match patterns and replicate the output of complex systems since an audio signal's degradation and consequential MOS score follows a pattern in relation to the impairment factors (choice of codec, delay, packet loss, jitter) of a system or network. Many different approaches utilising ANNs have been proposed to measure VoIP and, more recently, VoLTE voice quality. [14] demonstrated an ANN's ability to predict call quality based on packet loss, codec and talker identity. This was improved upon by [15] who developed a Random Neural Network to assess VoIP quality using the factors of delay, packet-loss, jitter and codec which, when compared to PESQ, proved to be almost as accurate. [21] built on this study by assessing VoLTE call quality using these metrics, however, their study was limited to the AMR-WB codec which was the only widely-used VoLTE codec at the time, again achieving scores closely correlated to PESQ. More recently [22], using Deep Belief Networks, achieved an accuracy of 96.1% relative to PESQ – outperforming estimative standards such as the E-Model. Deep Feed-Forward Neural Networks have also been used to predict VoIP call quality using similar parameters and produced predictions showing a high correlation (0.8693) with the top intrusive measurement standard, POLQA, in [23]. However these methods still failed to outperform PESQ. No studies which examined CNNs for this use-case were discovered.

### C. Convolutional Neural Networks and Transfer Learning

Transfer learning allows developers to take advantage of the work invested in successful models such as VGG-16, ResNet50, Inception v3 and MobileNet, by incorporating them into the earlier layers of their model to initialize the weights and thresholds of the network. The part of the model that defines functionality is then added in the later layers to "fine-tune" the existing model for the required purpose. Many of the principals of supervised learning apply to transfer learning as it is essentially the same process but with a domain-shift added. This approach has been successfully utilised by models such as NimaNet [24], a perceptual model to score degraded images on quality without the need for a reference. While networks developed for computer-vision functions are usually the basis for transfer learning due to their huge sample datasets and years of training, this does not preclude their utility in other domains. Transfer learning extends to classification of audio

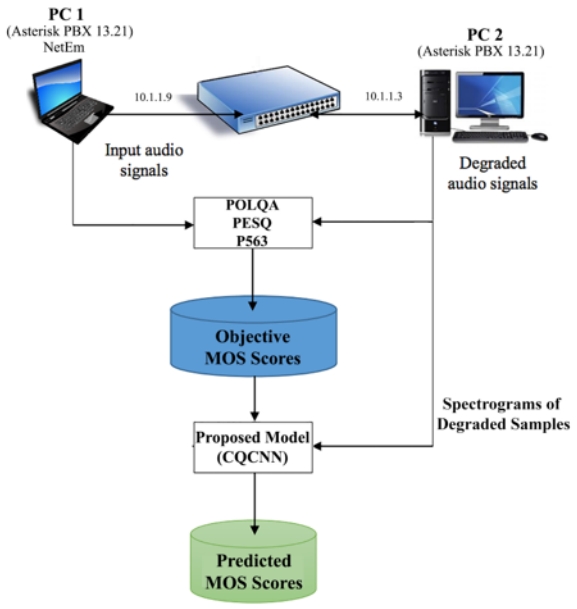


Fig. 1. Experimental architecture used.

data and increases the overall accuracy. Most audio applications use CNNs to evaluate spectrograms which graph a sound in 3 dimensions: time in the X axis; frequency in the Y axis and intensity of frequency represented by saturation/brightness – essentially the Z axis. This is done by taking a Short-Time-Fourier-Transform (STFT), which performs Fast Fourier Transform (FFT) operations at an interval or 'window' across the waveform. These windows and their overlap can be altered to highlight different features of the soundwave. FFTs are also used by the POLQA algorithm in computation of MOS scores [11] which suggests the potential utility of spectrograms in evaluating voice quality.

## III. Methodology

In this paper, we propose the use of a convolutional neural network as a novel, single-ended, objective method to evaluate call quality in VoLTE environments. The Call Quality Convolutional Neural Network (CQCNN) in this paper was trained using samples from the VoxForge [20] dataset using NVIDIA's CUDA GPU technologies for deep learning and implemented with python code in Jupyter notebooks. An overview of this project's process is illustrated in figure 1. This is detailed in the five stages that follow:

### A. Dataset Identification, Exploration and Processing

While high quality datasets recorded to ITU-T P.800 standards do exist [29], these are typically commercial products which are prohibitively expensive. In order to build a dataset for the purposes of VoLTE call quality assessment, a large corpus of recordings from a variety of different speakers would be required. This process would have to adhere to ITU-T standards and would need to amass ∼160 hours of

audio (68.000 10-second samples) to produce a dataset of the magnitude needed to train a deep neural network. The challenge of compiling this volume of data is compounded by a lack of support for AMR/EVS encoded audio in VoLTE simulation environments which could be used to efficiently generate the impaired signals needed to label our dataset using ITU-T's objective evaluation methods. Without an accurate VoLTE simulation to degrade each signal in a controlled and realistic environment, degraded signals would either need to be degraded using a real VoLTE network or a simplified network emulation. In light of these challenges, we elected to re-purpose the existing Voxforge dataset using network emulation. The VoxForge dataset is a crowdsourced speech corpus of 73,412 samples from 6,000 speakers recorded at 48kHz. This is consistent with the bandwidth supported by most VoLTE-enabled smartphone's microphones. The content of each sample also meets with the majority of the ESTI criteria for subjective and objective voice quality tests [25]. Due to POLQA's dual mode-set operation it was necessary to use the entire dataset twice to train the model for NB and SWB samples independently. Each sample must also be passed through a network emulation where it is subjected to a combination of impairment factors (codec, delay, jitter and packet loss) to produce a degraded counterpart. This is accomplished by using Asterisk, an open-source framework for VoIP applications, to generate and record calls across a network impaired with pre-set conditions using Linux's native Network Emulation software (NetEm). This results in 54 test conditions for samples to be recorded as 16-bit NB audio files. These consist of 27 combinations of 3 values of each network parameter (delay, jitter and packet loss) within normal operational limits for voice networks [4] duplicated for the two codecs, EVS and Opus. Due to bandwidth limitations in Asterisk's recording function and a lack of support for Opus and EVS in RTP capture software such as Wireshark, SWB samples could not be recovered from the emulation and were instead impaired with the packet loss values only for each codec using a function in its native application. All samples are then processed and scored by POLQA, PESQ and P.563 in turn, with labelled results generated and saved to separate files in .csv format.

### B. Extraction, Transformation and Loading of data (ETL)

During this stage, the audio samples are transformed into log-scaled spectrograms and stored in a dataframe for processing by the proposed model. Wideband and Narrowband spectrograms were generated for each sample by adjusting the window length and re-processing the datasets. During training, it was discovered that the spectrograms with a longer window length (NB) produced more accurate predictions for both the NB and SWB speech samples and these were subsequently used to evaluate the model. Each sample was normalized to 10 seconds in length by trimming or padding the spectrogram to produce a standardised input for the CQCNN model. While some samples exceeded this and lost data due to the trimming, 99.6% of the dataset was unaffected. The dataframe containing the spectrograms was then merged with the POLQA values which were to serve as the "ground-truth" labels when training and validating the model. This dataframe was then cleansed of any empty values which were the result of errors during POLQA's evaluation.

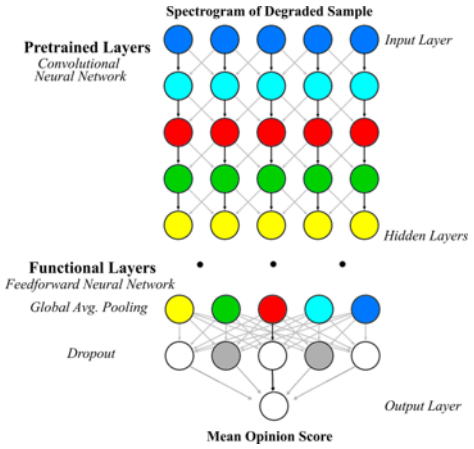


Fig. 2. Proposed Model (CQCNN) Architecture

### C. Model Definition

The proposed model's architecture is presented as a network diagram in figure 2. The CQCNN model uses a convolutional neural network based on the VGG-16 architecture. This was identified during initial tests as the best performing model over ResNet, Inception and VGG-19 architectures (table I). These layers of the network are pre-trained with weights from the famous ImageNet dataset to boost performance (table II). Global average pooling is then used to mimic the functionality of several fully connected layers, summarizing the features learned by the convolutional layers above. The final layer of this functional, fully-connected part of the architecture is a single node with a linear activation function which outputs a continuous value – the model's predicted MOS score for the sample. While CNNs typically operate as classification models, a regression-style model was better suited to the output values required for this use-case.

### D. Model Training, Validation and Optimization

The CQCNN model is then trained using 80% of the data and validated on the remaining 20%. A batch size of 20 was used to reduce training time while minimizing test error values. The adaptive learning rate optimizer ADAM was used to minimize mean squared error (MSE) during the training process. Mean absolute error (MAE) was also calculated to allow for simpler evaluation of the model. No more than 8-12 epochs were required to reach minimal test error values although training times of up to 30 epochs were tested. All training and modelling tasks were conducted using the Keras and Tensorflow libraries.

### E. Model Evaluation

The final stage of the project involved analyzing the model's predictions for the unseen 20% of samples. This involved the

use of python's pandas dataframes to compare and calculate various measures of each method's performance relative to the POLQA algorithm. In order to present an intuitive measure of the model's accuracy, mean relative error as accuracy is used rather than presenting the results in terms of error values such as MAE/MSE. This is defined by the following equation where $Y_T$ is the true (POLQA) score, $Y$ is the estimated value produced by the model and $n$ is the total number of samples:

$$MRE_{ACC} = \frac{100\%}{n} \sum_{i=1}^{n} \left( \frac{|Y - Y_T|}{(Y_{Tmax} - Y_{Tmin}) + Y_T} \right) \quad (4)$$

The inclusion of maximum and minimum $Y_T$ values ensures that all values are calculated relative to the range of possible POLQA MOS values, (1-4.75 for SWB samples and 1-4.5 for NB samples). All error values are therefore normalized and absolute ensuring that a percentage value between 1 and 100 is produced, regardless of the error's magnitude or bias.

## IV. Experimental Results and Analysis

The following results were collated by comparing the predictions generated from the proposed model (CQCNN) during the evaluation stage with MOS values obtained from ITU-T standardized objective methods for voice quality measurement (POLQA, PESQ and P.563).

TABLE I
PRETRAINED CNN RESULTS FOR VARIOUS ARCHITECTURES

| Architecture | MSE | MAE |
|---|---|---|
| VGG16 | 0.1274 | 0.2619 |
| VGG19 | 0.1623 | 0.3050 |
| ResNet50 | 0.1803 | 0.3640 |
| Inception v3 | 0.2063 | 0.3406 |

Table I shows that the VGG16 architecture produced the lowest loss/error during the training and validation process for the NB dataset. These results correlated when the SWB data was processed by the model. The VGG19 architecture was therefore selected as the candidate model for the rest of the evaluation stage.

TABLE II
EFFECTS OF PRE-TRAINING ON MODEL PERFORMANCE

| Weight Initialisation | MSE | MAE |
|---|---|---|
| Random | 0.1506 | 0.2968 |
| ImageNet | 0.1274 | 0.2619 |

Table II demonstrates the impact of pre-training on the best-performing model. These results confirmed that the pre-trained model did perform better than one initialised with random weight values, even if only by a small amount. The relatively low impact of pre-training in this case may be due to significant difference in dataset and use case compared to the VGG16 network that was trained on the imagenet weights. It was shown in [30] that transfer learning was more viable when the model is being utilised for similar purposes to the one its weights were trained for. In this case, the function differs significantly as does the dataset so transfer learning is of minimal use, but still boosts model performance.

TABLE III
MEAN RELATIVE ACCURACY VS. POLQA

| Method | Relative Accuracy (%) | |
|---|---|---|
| | ***NB*** | ***SWB*** |
| CQCNN | 91.93 | 82.50 |
| PESQ | 90.70 | 94.22 |
| P.563 | 86.09 | 62.36 |

TABLE IV
REPORTED ACCURACY OF RECENT DEEP CALL QUALITY EVALUATION MODELS

| Model | Reported NB Accuracy (%) |
|---|---|
| Yang et al. (2016)[23] | 86.80 |
| Affonso et al. (2018)[22] | 87.10* |
| Proposed Model (CQCNN) | 91.93 |

[a]Relative POLQA accuracy estimated from comparable PESQ results.

Table III demonstrates the proposed model's ability to predict the MOS values of both NB and SWB samples to a high degree of accuracy to those produced by the POLQA algorithm. The CQCNN model achieved an accuracy of 91.93% for the NB dataset, surpassing even the objective, intrusive PESQ algorithm. CQCNN's performance is shown to exceed the accuracy of the various models examined in section II (Table IV). The model also comfortably outperformed the only other single-ended, non-intrusive method (P.563) by 6%-20% in both the NB/SWB datasets respectively. Surprisingly, PESQ achieved a higher accuracy on the SWB dataset than CQCNN and even surpassed its own score from the NB dataset, despite not offering explicit support for SWB signal evaluation. The effect of quality, in terms of sample bandwidth, can be seen to have a significant impact on the proposed model's relative accuracy, with CQCNN performing 9.44% poorer while evaluating SWB samples. This may be due to sub-optimal window lengths set used during the production of spectrograms for the SWB dataset rather than a problem with the model itself, however, further work would be needed to determine the cause.

While CQCNN performed better on lower bandwidth samples, Table V shows that within each dataset it achieved 10%-20% greater accuracy in predicting the MOS values for high quality samples, where than value was greater than the midrange value for each dataset's range (2.75/2.875), than samples with a score below this threshold. Upon further analysis, a potential reason for this sizable discrepancy was identified.

TABLE V
PROPOSED MODEL (CQCNN) VS. SAMPLE QUALITY

| Condition | Relative Accuracy (%) | |
|---|---|---|
| | *CQCNN (NB)* | *CQCNN (SWB)* |
| Low Quality (MOS $<$ 2.75/2.875) | 80.72 | 59.29 |
| High Quality (MOS $>$ 2.75/2.875) | 92.76 | 88.75 |

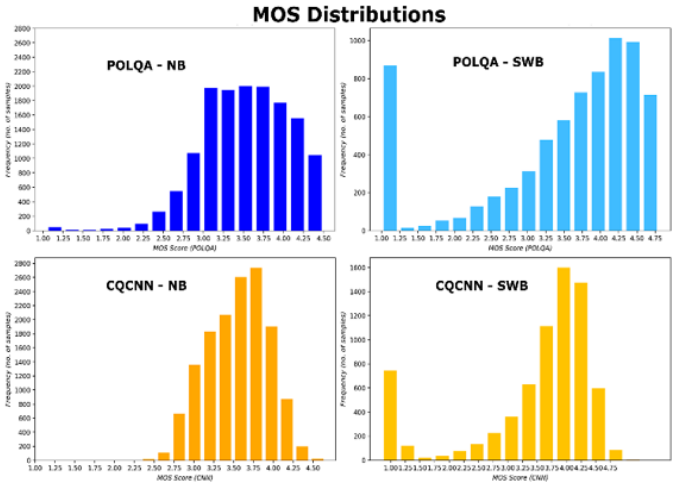


Fig. 3. Distribution of MOS scores for POLQA and CQCNN illustrating the number of samples (Y axis) with a given MOS score (X axis) for both SWB and NB datasets.

The distribution of MOS values produced by POLQA and CQCNN for each dataset used during validation and testing, shown in figure 3, illustrates a strong positive skew. With a very small number of samples in each dataset present below the midrange threshold, there is insufficient data for the model to learn the features that define samples at these scores. Boosting the number of low-quality samples in each dataset and retraining the model would potentially result in higher accuracy for the model's predictions. Figure 4 illustrates the mean MOS values produced for each of the 27 test conditions by the objective methods assessed in this paper. There is a clear correlation between the PESQ and POLQA algorithms and a greater variance in the mean MOS produced for each test condition by these objective-intrusive algorithms. The single-ended methods (P.563 and CQCNN) demonstrate mean MOS scores with little to no correlation with the test conditions. The poor reflection of different impairment conditions in these methods is understandable due to the exclusion of reference signals from their analysis. It is also possible that the spectrograms evaluated by the CQCNN do not represent these degradations effectively enough to be learned and associated with the various labelled test conditions. It is possible that the use of raw audio data as an input and perhaps the use of a Recurrent Neural Network to evaluate temporal features of the signals may yield a higher correlation between scores and network impairment factors and better performance overall.

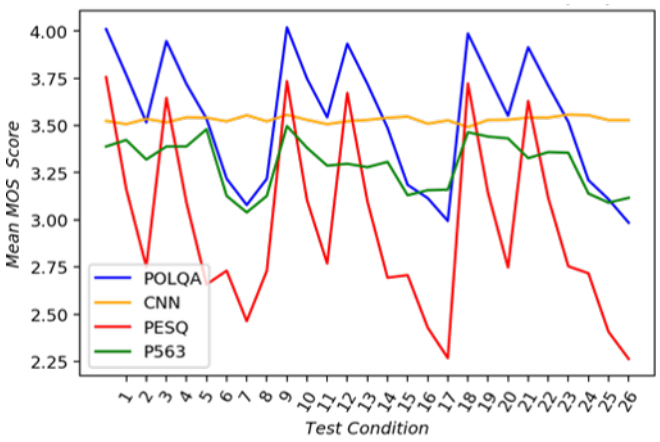


Fig. 4. Mean MOS score for each of the 27 NB test conditions, grouped by evaluation method.

## V. CONCLUSION

In this paper, we have utilized deep learning and Convolutional Neural Networks to produce a novel, single-ended, objective call quality evaluation method (CQCNN). This model was then trained on spectrograms of degraded speech produced through a high-quality, realistic simulation process by encoding samples of the VoxForge dataset with the latest AMR codecs EVS and Opus and passing them through a VoLTE network emulation impaired with adverse network conditions such as delay, jitter, packet-loss. In the results presented, the model achieved accuracies of 91.9% (NB) and 82.5% (SWB) in its predictions. This demonstrates that this model is capable of outperforming current objective standards for call quality evaluation, significantly outperforming the only standardized, non-intrusive method (P.563) when evaluating both NB and SWB and surpassing the accuracy of intrusive algorithms (PESQ) in predicting NB MOS values. CQCNN improves upon previous proposals for call quality evaluation systems that utilize deep learning both in terms of accuracy and scope. Our model demonstrates the ability to handle more complex, higher quality signals such as the SWB audio transmitted within VoLTE networks and presents the possibility for a single-ended quality evaluation tool that can perform well even in comparison to the high-accuracy referenced-based systems such as POLQA. With further refinements to the dataset and the methodology, it is possible to illict even higher levels of performance from the proposed model in the future. This could lead to the development of a significantly more available and practical implementation of call quality assessment that could be used within business, consumer and industry telecommunications networks to measure performance in real-time. The advent of 5G networks could allow for the implementation of such a model on consumer devices that could be extended using other well-researched deep learning methods such as generative adversarial models (GANs) or variational auto-encoders (VAEs) to perform audio super-resolution or signal interpolation so as to actually improve call quality, online and in real-time.

## ACKNOWLEDGMENT

The authors would like to express thanks to Rhode & Schwartz SwissQual AG for the use of their software implementation of POLQA.

## REFERENCES

[1] Karapantazis, S. & Pavlidou, F. (2009), "VoIP: A comprehensive survey on a promising technology", Computer networks (Amsterdam Netherlands), 53(12), pp. 2050-2090

[2] Rämö, A. and Toukomaa, H. (2015), "Subjective quality evaluation of the 3GPP EVS codec," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 5157-5161

[3] SPIRENT (2014), "IMS Architecture – The LTE User Equipment Perspective", White Paper, SPIRENT plc.

[4] Huawei (2018), "Vo5G Technical White Paper", Huawei Technologies Co. Ltd.

[5] Atti, V. et al. (2015), "Improved Error Resilience for VoLTE and VoIP with 3GPP EVS Channel Aware Coding", ICASSP'15, IEEE, pp.5713-5717

[6] Lecomte, J. et al. (2015), "Packet-loss Concealment Technology Advances in EVS". ICASSP'15, IEEE, pp.5708-5712

[7] IETF (2012), "RFC-6716: Definition of the Opus Codec", International Engineering Task Force, September 2012.

[8] Colins, D. (2003), "Carrier Grade Voice over IP", McGraw Hill Education.

[9] Daengsi, T., Wutiwiwatchai, C., Preechayasomboon, A. & Sukparungsee, S. (2014), "IP Telephony: Comparison of Subjective Assessment Methods for Voice Quality Evaluation", Walailak journal of science and technology, 11(2), pp. 87-92

[10] Luska, D., Fajt, S. and Krhen, M. (2014). "Sound quality assessment in VOIP environment.", 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE

[11] Beerends et al. (2013), "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I–Temporal Alignment", Journal of the Audio Engineering Society, 61(6), pp. 385-402.

[12] ITU-T (2018), "Recommendation P.862: PESQ", International Telecommunications Union.

[13] OPTICOM (2006), "P.563 Single-Sided Speech Quality Measure", OPTICOM, 3 March 2006

[14] Sun, L. and Ifeachor, E. (2006). "Voice quality prediction models and their application in VoIP networks.", IEEE Transactions on Multimedia, 8(4), pp.809-820.

[15] Radhakrishnan, K. and Larijani, H. (2011). "Evaluating perceived voice quality on packet networks using different random neural network architectures.", Performance Evaluation, 68(4), pp.347-360

[16] Goodfellow, I. (2016), "Deep Learning", MIT Press, Cambridge, MA.

[17] Van Den Oord, A. et al. (2016) "WaveNet: A Generative Model for Raw Audio", Proc. 9th ISCA Speech Synthesis Workshop, 2016. pp.125-125

[18] Abdel-Hamid, O. et al. (2012), "Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 4277-4280.

[19] Fisher, K.S., & Scherlis, A. (2016), "WaveMedic : Convolutional Neural Networks for Speech Audio Enhancement", Stanford University, Stanford, CA.

[20] VoxForge (2019), "VoxForge: Speech Corpus" - http://voxforge.org (Accessed 2 April 2019)

[21] Nguyen, Duy-Huy et al. (2016) Predicting VoLTE Quality using Random Neural Network. International Journal of Applied Information Systems. [Online] 11 (3), 1–5

[22] Affonso et al. (2018), "Speech Quality Assessment Over Lossy Transmission Channels Using Deep Belief Networks," in IEEE Signal Processing Letters, vol. 25, no. 1, pp. 70-74, Jan. 2018.

[23] Yang, H. et al. (2016), "Parametric-based non-intrusive speech quality assessment by deep neural network," 2016 IEEE International Conference on Digital Signal Processing (DSP), Beijing, 2016, pp. 99-103.

[24] Talebi, H. and Milanfar, P. (2018), "NIMA: Neural Image Assessment", Cornell University, 26 April 2018

[25] ETSI (2013), "Speech and Multimedia Transmission Quality (STQ); Speech Samples and their Usage for QoS Testing", ESTI TR 103 138 v.1.1.1, European Telecommunications Standards Institute, October 2013.

[26] E. Cipressi and M. L. Merani, "A Comparative Study on the Quality of Narrow-Band and Wide-Band AMR VoLTE Calls," 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 2019, pp. 1273-1278, doi: 10.1109/IWCMC.2019.8766598.

[27] Y. Han and G. Muntean, "Hybrid real-time quality assessment model for voice over IP," 2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Ghent, 2015, pp. 1-6, doi: 10.1109/BMSB.2015.7177263.

[28] E. Cipressi and M. L. Merani, "An Effective Machine Learning (ML) Approach to Quality Assessment of Voice over IP (VoIP) Calls," in IEEE Networking Letters, 2020, doi: 10.1109/LNET.2020.2984721.

[29] NTT-AT (2018), "Super Wideband Speech Database", NTT Advanced Technologies Corporation. Available at: http://www.ntt-at.com/product/widebandspeech/ (Accessed 12 May 2020)

[30] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? InThe IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.