Towards a Non-Intrusive Context-Aware Speech Quality Model

Rahul Jaiswal

Department of Information & Communication Technology
University of Agder, Grimstad, Norway
rahul.jaiswal@uia.no

Abstract

Understanding how humans judge perceived speech quality while interacting through Voice over Internet Protocol (VoIP) applications in real-time is essential to build a robust and accurate speech quality prediction model. Speech quality is degraded in the presence of background noise reducing the Quality of Experience (QoE). Speech enhancement algorithms can improve speech quality in noisy environments. The publicly available NOIZEUS speech corpus contains speech in environmental background noise babble, car, street, and train at two Signal-to-noise ratios (SNRs) 5dB and 10dB. Objective Speech Quality Metrics (OSQM) are used to monitor and measure speech quality for VoIP applications. This work proposes a Context-aware QoE prediction model, CAQoE, which classifies the speech signal context (i.e., noise type and SNR) in order to allow context-specific speech quality prediction. This work presents context-classification component of the proposed model. Speech enhancement algorithms are used in conjunction with an OSQM to estimate Mean Opinion Score (MOS) of noisy and enhanced samples to train machine learning classifiers to classify the speech signal context. For the different noise classes tested, a Decision tree classifier has better classification accuracy.

Index Terms: non-intrusive, speech quality, noise, speech enhancement, P.563, MOS, classifier, VAD, VoIP, QoE.

1. Introduction

With the development of wireless networks and growing popularity of mobile devices, adoption of VoIP is increasing. Real-time speech quality monitoring is essential to provide predictions of the actual speech quality experienced by the users of VoIP applications such as Google Meet, Microsoft Skype, and Apple FaceTime. As compared to the tedious and expensive subjective listening test i.e., Absolute Category Rating (ACR) [1] to measure speech quality, the objective speech quality assessment metrics are more practical and faster.

There are two categories of objective speech quality assessment methods, namely; intrusive and non-intrusive. Intrusive methods e.g. Perceptual Evaluation of Speech Quality (PESQ) [2], Perceptual Objective Listening Quality Assessment (POLQA) [3] and ViSQOL [4] are unsuitable for monitoring real-time speech quality because of the practicalities of accessing both the input reference signal and the received signal. Non-intrusive methods, such as ITU-T P.563 [5] estimate speech quality using only the received (degraded) signal. P.563 extracts the dominant distortion class and maps it to a single MOS score, which describes speech quality on a scale from 1 (bad) to 5 (excellent). An implementation of P.563 is publicly available. Parametric models e.g. the E-Model [6] are used to estimate speech quality using network parameters, e.g. network delay and packet loss and terminal parameters, e.g. jitter buffer overflow, coding distortion, and echo cancellation. However, they do not use the signal and thus are not suited to predict speech quality based on signal-noise characteristics [7].

For real-time speech quality monitoring, a real-time no-reference signal-based speech quality model is considered the most appropriate. No such signal-based, context-aware speech quality prediction models are described in the literature, motivating the proposed Context-aware QoE prediction model, "CAQoE". The proposed model could be deployed by the internet service providers to continuously monitor the service performance quality by detecting impairments and potentially identifying the context. The QoE-aware management actions could be then taken to maintain the user QoE levels [8].

2. Proposed Model

The outline of the proposed context-aware QoE model is shown in Fig. 1. The model comprises three main components: (i) a context-classifier that classifies the speech signal context (i.e., noise type and SNR); (ii) a Voice Activity Detector (VAD) to identify the voiced segments of noisy signal; and (iii) Speech Quality Model (SQM) which is a collection of noise specific neural networks trained to evaluate speech quality under specific noise conditions. The model is based on the hypothesis that by classifying the noise-type and intensity of the signal being evaluated can be routed to a quality assessment model that has been trained and optimised to a particular degradation.

2.1. Context-Classifier

The present work is focused on the context-classifier. The context-classifier aims to make use of the observation that different speech enhancement algorithms perform better in different contexts i.e., with different noise types and SNRs. Each speech enhancement algorithm uses a different noise estimation algorithms to separate the target speech with varying success. Using this knowledge, the model takes the input noisy signal and processes it using 12 standard speech enhancement algorithms [9, 10]. Along with the original unprocessed input signal, we now have 13 variations including the original signal. These signals are then processed with the objective speech quality metric (P.563) [5] to output 13 quality predictions (MOS) that are combined as an input feature vector to a ML classifier.

We anticipated that the classifiers would be able to learn from the relationship between the unprocessed signal quality estimates and the enhanced speech quality estimates in order to correctly classify the noise type and SNR (context). Seven classifiers, namely; XGBoost (XGB), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), K-nearest Neighbors (KNN) and Linear SVC (Lin. SVC), are investigated to identify the most appropriate ML classifier component of the proposed model.

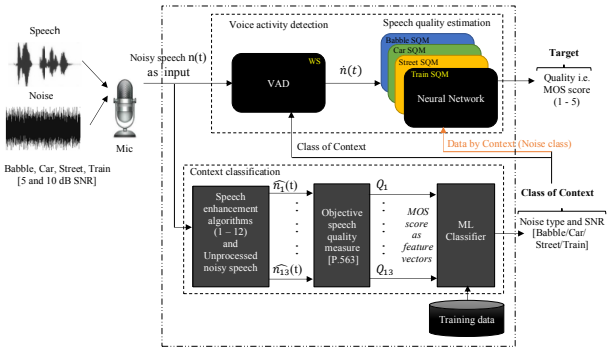

Figure 1: *Block diagram of proposed Context-Aware QoE Model.*

2.2. Voice Activity Detection

The second component in the proposed model is a voice activity detector (VAD). It identifies the voiced segments of the noisy signal based on the speech feature (spectral centroid) [11]. It is incorporated into the overall CAQoE model as a pre-processing stage prior to presenting the signal to the speech quality neural network component. The VAD used here is based on our previous work [11] which found that a weighted spectral centroid VAD is preferred for speech data with environmental noise.

2.3. Speech Quality Estimation

The third component in the proposed model is used to compute an estimate of speech quality and its performance evaluation. This component contains a collection of Deep Neural Networks (DNNs), each trained and optimised for a particular noise class. The classifier activates the chosen DNN which is then used to evaluate the noisy signal and output a target value of the predicted speech quality (MOS) for the noisy signal.

3. Evaluation of the Classifier

The context-classifier takes 30 noisy samples of each noise types: babble, car, street and train at two SNRs (5dB and 10dB) from the NOIZEUS corpus [9]. The total number of samples is 240 (30 samples x 4 noise types x 2 SNRs). Each noisy sample is processed with 12 speech enhancement algorithms and the unprocessed sample. MOS scores are estimated for these 13 conditions using the P.563 metric. The MOS quality labels are used as an input feature vector to train a ML classifier to classify the speech signal context (i.e., the noise type and SNR).

We have 30 samples in each class (where a class refers to a noise type and SNR combination e.g., Babble 5dB). This resulted in a small amount of data with which training a classifier gave a poor classification accuracy of 35 % for an eight class classifier. Therefore, we chose to perform binary classification with imbalanced datasets/classes. Out of eight classes, we assigned the first class as "class 0" and the remaining seven classes as "class 1", and labelled it as, e.g. "Babble 5dB". Similarly, we assigned the second class as "class 0" and the remaining seven classes as "class 1", and labelled it as "Babble 10dB". We followed the same strategy for the remaining classes. To balance each class, we reduced the size of majority class data (class 1) equals to minority class data (class 0) using Synthetic Minority Oversampling Technique (SMOTE) [12]. Data of each class is divided into 80:20 ratio for training and testing. We used 10-fold cross validation technique to validate the effectiveness of ML classifiers and to prevent over-fitting.

For imbalanced class distribution, F-measure and Geometric mean (G-mean) or balanced accuracy are used to measure the classification performance [13]. F-measure is the measure of test's accuracy, and is defined as the weighted harmonic mean of precision and recall. G-mean is the geometric mean of the classification precision of minority class and the classification precision of majority class. It evaluates the model's ability to correctly classify the minority and the majority class [14].

4. Results and Discussion

It can be seen from Table 1 and Table 2 that the XGBoost (XGB) and the Decision Tree (DT) have highest test accuracy (92 %) for babble 5dB among all classifiers. The DT has highest accuracy for street 5dB, babble 10dB, car 10dB and train 10dB noise class whereas the LR and linear SVC has the highest accuracy for car 5dB and train 5dB noise class. The Linear SVC exhibits better accuracy for street 10dB noise class. In most of the cases, 5dB SNR noise classes have better classification accuracy as compared to 10dB SNR noise classes. Average test accuracy of the DT is 77 %, which is the highest among all ML classifiers in classifying the speech signal context.

Table 1: *F-measure of each classifier for each class*

Classifier→ Class ↓	XGB	RF	DT	LR	SVM	KNN	Lin. SVC
Babble 5dB	0.92	0.86	0.92	0.71	0.73	0.83	0.62
Babble 10dB	0.77	0.67	0.83	0.71	0.60	0.57	0.62
Car 5dB	0.67	0.71	0.67	0.80	0.40	0.67	0.80
Car 10dB	0.80	0.67	0.83	0.73	0.67	0.67	0.73
Street 5dB	0.83	0.77	0.91	0.77	0.60	0.83	0.77
Street 10dB	0.67	0.44	0.57	0.77	0.44	0.25	0.83
Train 5dB	0.67	0.67	0.67	0.83	0.71	0.40	0.83
Train 10dB	0.44	0.46	0.80	0.57	0.60	0.53	0.62
Average	0.72	0.65	0.77	0.73	0.59	0.59	0.72

Table 2: *G-Mean of each classifier for each class*

Classifier→ Class ↓	XGB	RF	DT	LR	SVM	KNN	Lin. SVC
Babble 5dB	0.91	0.81	0.91	0.64	0.74	0.83	0.57
Babble 10dB	0.74	0.66	0.83	0.64	0.64	0.47	0.57
Car 5dB	0.66	0.64	0.66	0.81	0.47	0.66	0.81
Car 10dB	0.81	0.70	0.83	0.74	0.66	0.66	0.74
Street 5dB	0.83	0.74	0.91	0.74	0.64	0.83	0.74
Street 10dB	0.66	0.52	0.47	0.74	0.52	0.37	0.83
Train 5dB	0.70	0.66	0.70	0.83	0.64	0.47	0.83
Train 10dB	0.52	0.40	0.70	0.47	0.64	0.33	0.57
Average	0.72	0.64	0.75	0.70	0.61	0.57	0.70

5. Future Work

The ongoing published work [15] has presented the context-classifier component of the proposed model (CAQoE) for classifying the speech signal context (i.e., noise type and SNR). Future work will develop a collection of DNNs, trained and optimised for particular noise classes i.e., context-sensitive. It will also extend the range on contexts for the context-classifier.

6. Acknowledgement

I would like to thank my advisor Dr. Andrew Hines for his constant guidance and invaluable discussions.

7. References

[1] "ITU-T Recommendation P.800: Methods for subjective determination of transmission quality," 1996.

[2] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)- a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, pp. 749–752.

[3] "ITU-T Recommendation P.863: Perceptual Objective Listening Quality Assessment (POLQA)," *International Telecommunication Union, Geneva, Switzerland*, 2011.

[4] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[5] "ITU-T Recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications," *International Telecommunication Union, Geneva, Switzerland*, 2004.

[6] J. A. Bergstra and C. Middelburg, "ITU-T Recommendation G. 107: The E-Model, a computational model for use in transmission planning," 2003.

[7] S. Möller, W. Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

[8] H. Z. Jahromi, A. Hines, and D. T. Delanev, "Towards Application-Aware Networking: ML-Based End-to-End Application KPI/QoE Metrics Characterization in SDN," in *Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2018, pp. 126–131.

[9] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 1, 2006, pp. 153–156.

[10] ——, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[11] R. Jaiswal and A. Hines, "The Sound of Silence: How Traditional and Deep Learning Based Voice Activity Detection Influences Speech Quality Monitoring," in *AICS*, 2018, pp. 174–185.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence research*, vol. 16, pp. 321–357, 2002.

[13] L. Der-Chiang, S. C. Hu, L.-S. Lin, and C.-W. Yeh, "Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced datasets," *PLOS ONE*, vol. 12, no. 8, pp. 1–24, 2017.

[14] S. Belarouci and M. A. Chikh, "Medical imbalanced data classification," *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 3, pp. 116–124, 2017.

[15] R. Jaiswal and A. Hines, "Towards a Non-Intrusive Context-Aware Speech Quality Model," in *31st Irish Signals and Systems Conference (ISSC)*. IEEE, 2020, pp. 1–5.