The sound of silence: how traditional and deep learning based voice activity detection influences speech quality monitoring

Rahul Jaiswal, Andrew Hines

School of Computer Science, University College Dublin, Ireland
rahul.jaiswal@ucdconnect.ie, andrew.hines@ucd.ie

Abstract. Real-time speech quality assessment is important for VoIP applications such as Google Hangouts, Microsoft Skype, and Apple FaceTime. Conventionally, subjective listening tests are used to quantify speech quality but are impractical for real-time monitoring scenarios. Objective speech quality assessment metrics can predict human judgement of perceived speech quality. Originally designed for narrow-band telephony applications, ITU-T P.563 is a single-ended or non-intrusive speech quality assessment that predicts speech quality without access to a reference signal. This paper investigates the suitability of P.563 in Voice over Internet Protocol (VoIP) scenarios and specifically the influence of silences on the predicted speech quality. The performance of P.563 was evaluated using TCD-VoIP dataset, containing speech with degradations commonly experienced with VoIP. The predictive capability of P.563 was established by comparing with subjective listening test results. The effect of pre-processing the signal to remove silences using Voice Activity Detection (VAD) was evaluated for five acoustic feature-based VAD algorithms: energy, energy and spectral centroid, Mahalanobis distance, weighted energy, weighted spectral centroid and four Deep learning model-based VAD algorithms: Deep Neural Network, Boosted Deep Neural Network, Long Short-Term Memory and Adaptive context attention model. Analysis shows P.563 prediction accuracy improves for different speech conditions of VoIP when the silences were removed by a VAD. The improvements varied with input content highlighting a potential to switch the VAD used based on the input to create a content aware speech quality monitoring system.

Index Terms: Speech Quality, VoIP, P.563, Voice Activity Detection (VAD)

1 Introduction

The availability of high speed mobile and fixed broadband connection is driving the adoption of Voice over Internet protocol (VoIP), which exploits packet-switching techniques and become both an alternative and a replacement for Public switched networks (PSTN) [2]. Real-time monitoring is essential to provide predictions of the actual speech quality experienced by users of the VoIP

applications such as Google Hangouts, Microsoft Skype, and Apple FaceTime. Speech quality is measured using subjective testing such as Absolute Category Rating (ACR) [1] and considered the most reliable method, but it is time consuming, impractical and inappropriate for real-time quality assessment. As an alternative approach, objective metrics are useful to the application developers and network system operators to ensure that the changes to the platform do not have a negative impact on user's quality of experience (QoE). This technique is faster, practical and appropriate for real-time quality assessment.

Single-ended objective metrics also referred to as no-reference metrics predict speech quality based on an evaluation of the degraded signal only, and could be deployed at the end point of VoIP channels to monitor quality [14]. Moller [14] reports that two non-intrusive estimation of narrow-band speech quality algorithms are standardized, the International Telecommunication Union (ITU) recommended P.563 [16] and the second, American National Standard Institute (ANSI) standard, referred to as 'Auditory non-intrusive quality estimation plus' (ANIQUE+) [10]. An implementation of P.563 is publicly available while ANIQUE+ is commercially available. Other examples of non-standardized speech quality prediction metrics are Low complexity speech quality assessment (LCQA) [3] and a recently presented DNN-based quality prediction metric [15]. Reported results indicated that the DNN-based metric performed poorly in competing speaker scenarios. This paper evaluates the ITU standard P.563 [16] no-reference metric to investigate the impact of silence in real-time speech quality monitoring. In a real-time monitoring scenario, longer silences than those expected by P.563 will distort the quality prediction. For example, P.563 specifies 25-75% speech activity as a valid input. If speech activity in a sample is less than 25%, the sample quality will be predicted as bad, as the input is invalid.

Early techniques used characteristics of the acoustic wave such as energy, spectral centroid, spectral entropy and zero-crossing rate to recognise the human voice based on threshold. These techniques first make assumptions on the distribution of speech and background noises (usually in the spectral domain), and then design statistical algorithms to dynamically estimate the model parameters, making them flexible in dealing with non-stationary noises [25]. But these methods have limitations e.g. model assumptions may not fully capture data distributions due to few parameters and they may not be flexible enough in fusing multiple acoustic features, which results in partial utilization of information in speech. Recently, data driven deep learning approaches have been adopted which have the potential to overcome the limitations of those model-based methods. The deep learning-based VADs can be integrated to the speech quality prediction systems to improve their performance. Also, they can fuse multiple features much better than traditional VADs e.g. deep neural networks (DNN), long short-term memory (LSTM). These models extract acoustic features during the course of training and uses different classifiers to classify speech and non-speech.

In this paper the suitability of applying an acoustic feature-based VAD and Deep learning model-based VAD to remove silences before presentation to P.563

is investigated. The paper is laid out as follows: Section 2 introduces the ITU no-reference objective metric P.563, Voice activity detectors (VADs), TCD-VoIP database, and Section 3 describes the evaluation method. Section 4 presents and discusses the results before concluding remarks and future directions are made in section 5.

2 Background

2.1 Single-ended Speech Quality Model: P.563

Single-ended, non-intrusive speech quality models are sometimes referred to as no-reference models. They are useful for real-time monitoring of speech quality and scenarios where a reference source is unavailable. One such model P.563 [13] was standardized by ITU in may 2004. It was designed for assessing samples of active speech and is used for narrow-band speech signals. The P.563 model is based on three principles [13] namely: a physical model of vocal tract; a reconstruction of the reference signal to assess unmasked distortions; and focusing on specific distortions: e.g. temporal clipping, robotization, and noise. Quality prediction by P.563 involves several stages: pre-processing, dominant distortion classification, and perceptual mapping. The degraded signal is pre-processed which involves reverse filtering, speech level adjustment, identifying speech portions and calculating speech and noise levels via a VAD [16]. Distortions classes include unnaturalness of speech, robotic voice, beeps, background noise, signal to noise ratio (SNR), mutes, interruptions are extracted from the voiced signal parts. A dominant distortion class is determined and mapped to a single mean opinion score (MOS) which describes speech quality on a scale from 1 (bad) to 5 (excellent). The pre-processing stage in P.563 includes a VAD which is based on adaptive power threshold, using an iterative approach. The internal VAD is used to compute a range of features and the active and inactive portions of speech activity as shown in Fig. 1 and does not pre-process to remove silence which can be seen in full details [16]. The performance of P.563 has been shown to be poor on samples containing silences [7]. The importance of silence detection is investigated here, using two VADs in system, the first VAD removes silences and the second VAD (inside P.563) is estimating speech activity.

2.2 Voice Activity Detector

Applications like speech and speaker recognition, speech enhancement, speech coding and speech synthesis need efficient feature extraction techniques where most of the voiced part contains speech attributes. Silence removal is a technique which is adopted for this purpose that facilitates the system to be computationally more efficient. The events in speech are classified as (*i*) silence; (*ii*) unvoiced; and (*iii*) voiced.

VAD detects presence and absence of human speech like silence, voiced and unvoiced based on speech features. It avoids unnecessary coding and transmis-

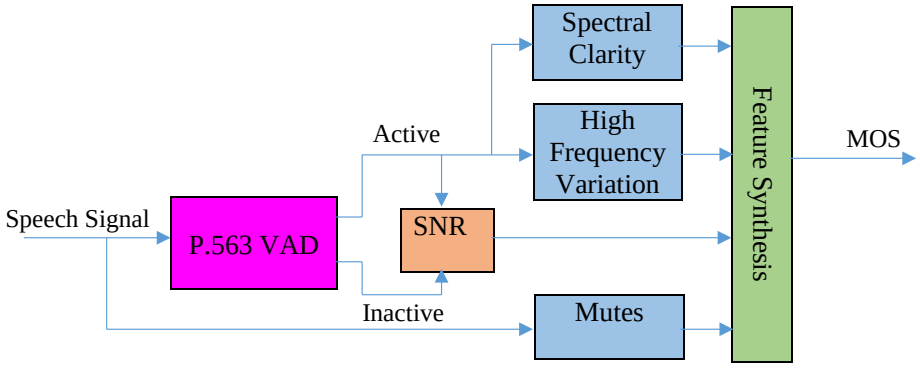

Fig. 1: VAD in P.563 [16]

sion of silence packets in Voice over Internet Protocol (VoIP) applications, saving computation and network bandwidth. For continuous monitoring of conversation, an aggressive VAD may yield better predictive performance from the P.563 metric [13]. Five different acoustic feature-based VAD algorithms and four Deep-learning model-based VAD algorithms are presented below and evaluated. **Ground Truth (GT):** This is the ideal binary mask that a VAD would create. It was computed using the known speech-silence structure of the speech samples in TCD-VoIP dataset (See Table 1). This mask is created as a representation of a VAD with perfect voiced activity detection. The speech samples are divided into overlapping 25 ms frame size with 10 ms shift.

Table 1: Format of Speech samples in TCD-VoIP [6]

Silence (1s)	Sentence 1	Silence (2s)	Sentence 2	Silence (1s)

Energy (E) [21]: The simplest method of voice activity detection is analysis of short time energy E of speech samples $x_i(n)$, given by equation (1) for each i^{th} frame having samples n, with frame length N. The silences are identified by finding frames with maximum energy (peak-to-peak) less than 0.003 joule.

$$E(i) = \frac{1}{N}\sum_{n=1}^{N}|x_i(n)|^2 \tag{1}$$

Energy and Spectral Centroid (ES) [5]: The speech samples are broken into overlapping frames of size 25 ms with 10 ms shift and for each frame, two speech features namely; short time energy, given by equation (1) and spectral centroid, given by equation (2) are extracted. The spectral centroid, C_i, of i^{th} frame is defined as the centre of gravity of its spectrum, given by equation (2),

$$C_i = \frac{\sum_{k=1}^{N}(k+1)X_i(k)}{\sum_{k=1}^{N} X_i(k)} \tag{2}$$

where $X_i(k), k = 1, 2, ...N$, is Discrete Fourier Transform (DFT) coefficients of i^{th} short time frame of frame length N. For each speech features, a simple threshold is calculated separately. As long as the two thresholds are estimated, a common threshold is computed and applied to extract the voiced segments [5].

Uni-dimensional Mahalanobis Distance (UMD) [19]: This VAD detects voiced segments from the speech samples using uni-dimensional Mahalanobis distance function which is a linear pattern classifier [22]. It assumes that background noise present in the utterances are Gaussian in nature and uses statistical properties (mean and standard deviation) of background noise to make a sample as voiced or silence. It considers that for a 1D Gaussian distribution, 99.7 % of its probability mass is in the range of mean $|\mu| \leq 3$ i.e. $Pr[|x - \mu| \leq 3] = 0.997$.

The speech samples are divided into overlapping frames of size 25 ms with 10 ms shift and the Mahalanobis distance [19] "r" from random variable x to mean μ measured in units of standard deviation σ is defined by equation (3). The steps of algorithm are detailed in [19].

$$r = \frac{|x - \mu|}{\sigma} \tag{3}$$

Weighted Energy and Weighted Spectral Centroid (WE) [5]: The energy and spectral centroid VAD has two speech features namely; short time energy and short time spectral centroid and the threshold is calculated as the weighted average between the two first local maxima of histogram for each feature individually [5]. Therefore, this VAD has been broken in two sub-parts, namely; weighted energy VAD and weighted centroid VAD.

Deep Neural Network (DNN) [23]: This DNN-based VAD implementation uses two classifier classes, namely: speech class and silence class. The input vector of the DNN is constructed for every frame of the speech signal based on multi-resolution cochleagram (MRCG) feature. MRCG is a concatenation of four cochleagram features with different window sizes and frame lengths. The cochleagram features are generated from gammatone filterbanks. The output vector is the posterior probabilities of these two classes, which is computed for each input vector based on threshold respectively. A DNN model is trained for an input feature sequence $x_m = [x_1, x_2, ...x_M]$ to an output sequence $y_m = [y_1, y_2, ...y_M]$, where $m = 1, ..., M$ is the time of frame. Cross-entropy criterion is used to optimize the VAD.

Boosted Deep Neural Network (bDNN) [26]: A Boosted DNN VAD that is an extended version of the DNN VAD. The main difference between training bDNN and DNN VAD is that for bDNN VAD, each speech frame is further expanded i.e. x_m to $x_m = [x_{m-W}, x_{m-W+1}, ..., x_m, ..., x_{m+W-1}, x_{m+W}]$ and similar for y_m, where $m = 1, ..., M$ is time of frame and "W" is user defined half-window size [26]. It has $(2W+1)d$ input and $2W+1$ output feature sequence.

Long Short-Term Memory (LSTM) [24]: The LSTM VAD architecture [20] is unidirectional with "k" hidden layers and "n" hidden units per layer. The input vector to the LSTM is a single 40-dimensional multi-resolution cochleagram (MRCG) features. The output vector is the posterior probabilities of silence and speech class, which is computed for each input vector based on threshold respectively. Cross-entropy is used as the optimisation criterion and truncated back propagation through time (BPTT) learning algorithm as the minimisation criteria [24].

Adaptive context attention model (ACAM) [11]: An ACAM-based VAD that is a frame-based speech or silence classifier. The input speech signal is divided into overlapping 25 ms frames with 10 ms shifts. 768-dimensional multi-resolution cochleagram (MRCG) feature vectors are extracted and trained using an algorithm called Adam for stochastic optimization [12].

2.3 Experimental Corpus

In order to investigate the influence of silence in VoIP applications, a dataset with silences and speech samples was used. TCD-VoIP [6] is a publicly available dataset of degraded speech samples with corresponding subjective opinion scores, generated by ACR test described in ITU-T Rec. P.800 [1]. Table 1 shows the format of speech samples. Each sample has a leading and trailing silence with 2 sec silence in middle (50% silences). The average length of each sentence is 2.4 sec. The speech samples are 16 bit WAV files, sampled at 48 kHz. The utterances have been taken from Harvard test sentence list. There were 24 listeners in all experiments and each condition was tested with 4 speakers (2 male and 2 female). It contains five types of platform independent degradation conditions (independent on codec, network or hardware), common to VoIP. These are: background noise, intelligible competing speakers, echo effects, amplitude clipping, and choppy speech. The competing speakers is a separate case of large-crowed babble as the speaker is intelligible and this is a common VoIP call scenario. The echo effects in a voice call usually occur due to transmitted speech being picked up in the receiving unit's microphone, creating a feedback loop like the scenario of microphone misplacement. Clipping occurs, when the amplitude of samples in a signal is set above the maximum permitted value e.g. microphone loudness. Choppy speech refers to speech, which is affected by missing samples due to packet loss in the VoIP network. The background noise sourced the AURORA database, developed by Hirsch et al. [9]. The summary of conditions and parameters are presented in [6].

3 Evaluation Method

The VAD mask, which is the binary mask, has been used to get the percentage of voiced segments of each VADs and compared to the ground truth (GT) mask to get the true positive/negative and false positive/negative which measures the performance of a particular VAD in terms of precision, recall and F-score [17].

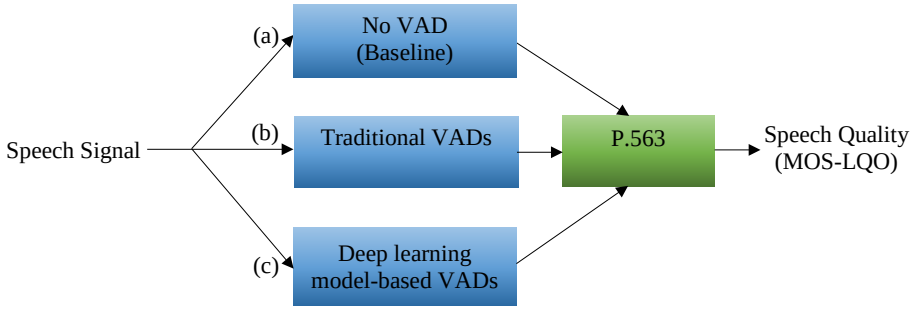

Fig. 2: Structure of experimental implementation with 3 paths (a) No VAD [Baseline]; (b) Traditional VADs (E, ES, UMD, WE and WS); and (c) Deep learning model-based VADs (DNN, bDNN, LSTM and ACAM).

F-score measures the accuracy and it's maximum value is 1 (precision = recall = 1) which means that voice activity decision of a certain VAD algorithm is equal to the reference transcription. The performance of a VAD deteriorates when the portion of active speech is mis-classified as non-active giving temporal clipping [8].

The pre-processed samples of TCD-VoIP via each VADs are used as the input to P.563 to get the objective predictions. To investigate the performance of P.563 in measuring speech quality, the per-condition objective predictions (MOS-LQO) of pre-processed samples are calculated to compare with the per-condition subjective mean opinion scores of samples (MOS-LQS). The degraded samples are re-sampled from 48 kHz to 8 kHz for testing. The test evaluates 384 speech samples. For each condition, 4 samples (2 male and 2 female speakers) are tested giving 96 conditions. The samples are processed to remove the WAV headers and evaluates with P.563 as a baseline. The Deep learning model-based VADs were trained using speech samples from the TIMIT corpus [4]. The TIMIT utterances have considerably shorter silences than speech which could create a class imbalance problem. To address this, the model developers added 2 second silence segments were before and after each utterance [11]. Multi-resolution cochleagram (MRCG) features were extracted as the input vector to these VADs. TCD-VoIP dataset are used as the test dataset. The samples are pre-processed by the VADs described in Section 2.2 to remove the silences and the resulting signals are evaluated with P.563 as illustrated in Fig. 2. The performance comparison between subjective listener test and objective metric prediction quality scores are quantified in terms of Pearson correlation coefficients (ρ_p), Spearman rank order coefficients (ρ_s) and root mean squared error ($RMSE$) [18]. Scatter plots of results classified by degradation condition are presented to visualise the statistics.

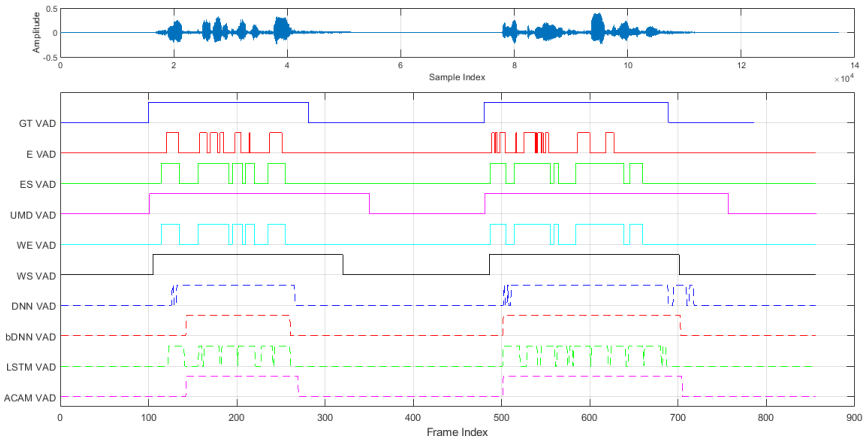

Fig. 3: Original signal, Traditional VADs mask and Deep learning model-based VADs mask compared to ground truth (GT) mask.

4 Results and Discussion

Fig. 3 shows a single speech sample, `C_16_ECHO_FG.wav`. The signal and computed VAD masks are presented. Masks for the traditional VADs and the data driven model-based VADs are compared to the ground truth (GT) mask. It can be seen from Fig. 3 that E, ES, WE, DNN, bDNN, LSTM and ACAM VAD are under-estimating the percentage of voice detected as compared to GT VAD. UMD VAD is over-estimating and WS VAD provides the best estimate compared to the ground truth (GT) mask. The performance has been measured on all 384 test samples of the TCD-VoIP dataset. The average percentage of voice detected of each degraded class with all VADs are compared to ground truth (GT) VAD and tabulated in Table 2. The Precision, Recall and F-scores have been calculated and the F-scores are presented in Table 3. The F-scores have been calculated for all degraded as well as all reference speech samples to examine and compare the accuracy of each VADs with reference transcription. An analysis of Table 2 shows that ES and WE VAD have same results except for noise conditions and both are under-estimating voiced segments as compared to GT VAD. The simple energy (E) VAD did not perform well. UMD VAD over-estimates voice for echo and competing speakers and under-estimates noise. It works well for the chop and clip conditions. WS VAD over-estimates voiced segments for noise. Similarly, all deep learning model-based VADs are under-estimating percentage of voiced segments with respect to GT VAD. Table 3 shows that UMD VAD has high F-score for echo and chop conditions but WS VAD has higher accuracy among all VADs for competing speakers, clip and all files. Boosted DNN VAD has higher accuracy for noise among all VADs and it is the best among all other deep learning VADs. WS VAD exhibited the best performance, when used as a pre-processing unit to P.563.

Table 2: Average percentage of voice detected in test samples.

VAD	Echo	Chop	Clip	Compspkr	Noise	All
GT	51.93	51.88	51.78	57.05	51.89	52.63
E	12.38	12.54	15.85	14.11	20.83	15.29
ES	28.48	27.72	32.58	29.86	31.66	29.88
UMD	55.19	51.00	51.43	71.05	33.55	50.50
WE	28.48	27.72	32.59	29.86	35.39	30.82
WS	44.90	41.21	44.54	49.12	56.71	47.50
DNN	39.85	36.26	35.92	44.98	49.23	41.47
bDNN	40.72	39.70	39.85	54.04	47.08	43.87
LSTM	47.86	49.37	48.15	54.08	45.47	48.59
ACAM	40.11	38.73	38.81	48.30	48.70	42.92

Table 3: F-scores by condition compared to ground truth (GT).

VAD	Echo	Chop	Clip	Compspkr	Noise	All (deg)	All (ref)
GT	1	1	1	1	1	1	1
E	0.239	0.237	0.469	0.394	0.464	0.277	0.391
ES	0.428	0.422	0.772	0.682	0.705	0.451	0.694
UMD	0.553	0.615	0.818	0.777	0.575	0.438	0.557
WE	0.428	0.422	0.772	0.682	0.695	0.460	0.694
WS	0.548	0.530	0.923	0.875	0.812	0.585	0.888
DNN	0.511	0.506	0.820	0.853	0.855	0.547	0.840
bDNN	0.515	0.520	0.869	0.865	0.890	0.562	0.868
LSTM	0.538	0.553	0.718	0.743	0.802	0.552	0.744
ACAM	0.513	0.513	0.857	0.883	0.886	0.556	0.857

The statistics of TCD-VoIP dataset, broken down by condition type are presented in Table 4. It can be seen from Table 4 that the Pearson correlation of noise and all samples of WS VAD gives good correlation among all VADs and higher than No VAD (baseline). For clip conditions, UMD VAD and for chop conditions, boosted DNN VAD is good. WS VAD has strong Spearman correlation for noise conditions as compared to No VAD (baseline). LSTM VAD is good for competing speaker conditions and ES VAD for clip conditions. WS VAD has the lowest RMSE for noise conditions and little increase for other conditions with respect to baseline. WS VAD exhibited better performance than the other VADs tested, when used as a pre-processing unit to P.563 and has higher correlation with subjective MOS score i.e. estimating better speech quality.

To explore the impact using a VAD as a pre-processing unit to P.563, the *N*o VAD (baseline) and WS VAD plots for objective metric P.563 (denoted by MOS-LQO) on the TCD-VoIP dataset have been compared to subjective results (denoted by MOS-LQS) per-condition for each type of degradation in Fig. 4 (a) - (b). The improvement for the noise condition (denoted with ◀) in particular is visually evident.

Fig. 4 (c) - (d) show the best traditional and best deep learning-based VAD performance with correlation between subjective and objective speech quality. While comparing the best deep learning-based VAD with best traditional VAD,

Table 4: Pearson correlations, Spearman rank correlations and RMSE (per-condition) for each degradation class of TCD-VoIP with a grouped result for all conditions.

VAD	ECHO			CHOP			CLIP		
	ρ_p	ρ_s	RMSE	ρ_p	ρ_s	RMSE	ρ_p	ρ_s	RMSE
No VAD	0.500	0.346	1.290	0.596	0.530	**0.720**	0.846	0.692	**0.600**
E	**0.854**	**0.818**	**0.642**	0.738	**0.777**	0.725	0.292	0.402	1.017
ES	0.314	0.183	1.757	0.737	0.640	1.275	**0.878**	**0.841**	1.062
UMD	0.663	0.624	1.193	0.437	0.429	1.035	0.808	0.692	0.884
WE	0.314	0.183	1.757	0.735	0.639	1.272	0.876	0.841	1.064
WS	0.606	0.449	1.302	0.643	0.613	1.006	0.842	0.771	0.923
DNN	0.600	0.464	1.467	0.672	0.612	1.176	0.459	0.353	1.229
bDNN	0.547	0.506	1.429	**0.765**	0.749	1.052	0.775	0.740	0.962
LSTM	0.417	0.292	1.307	0.534	0.469	0.819	0.747	0.586	0.693
ACAM	0.596	0.486	1.453	0.669	0.639	1.128	0.757	0.767	1.091

VAD	COMPSPKR			NOISE			ALL		
	ρ_p	ρ_s	RMSE	ρ_p	ρ_s	RMSE	ρ_p	ρ_s	RMSE
No VAD	0.675	0.607	**0.685**	0.750	0.719	0.758	0.589	0.523	**0.861**
E	0.402	0.466	1.034	0.593	0.654	0.949	0.637	**0.672**	0.865
ES	0.522	0.568	1.352	0.752	0.675	1.038	0.580	0.534	1.323
UMD	0.601	0.718	0.799	0.178	0.016	1.382	0.468	0.446	1.108
WE	0.522	0.568	1.352	0.778	0.674	1.037	0.591	0.534	1.322
WS	0.821	0.757	0.884	**0.841**	**0.774**	**0.703**	**0.665**	0.614	0.984
DNN	0.515	0.433	1.091	0.731	0.687	0.968	0.585	0.514	1.193
bDNN	0.782	0.735	0.927	0.766	0.683	0.843	0.663	0.611	1.079
LSTM	**0.846**	**0.837**	0.774	0.778	0.743	0.849	0.600	0.563	0.928
ACAM	0.829	0.700	0.937	0.649	0.546	0.967	0.601	0.565	1.137

it can be seen that bDNN VAD has a better f-score than weighted spectral centroid (WS) VAD for the CHOP speech condition in particular with little difference for other degradation conditions.

5 Conclusions and Future Directions

P.563 was designed as an objective no-reference metric but provides inaccurate predictions when input samples contain long silence segments. This paper reports benchmarking results of nine different voice activity detectors used as a pre-processing unit to P.563 tested using the TCD-VoIP database. The results show that for the VoIP conditions tested, a weighted spectral centroid pre-processing VAD improves prediction of speech quality for all conditions tested compared to baseline. The boosted DNN VAD exhibited better performance for the CHOP condition in particular. Pre-processing with a VAD demonstrated that P.563 could be deployed for real-time monitoring VoIP speech quality even in situations where there are longer periods without voice activity. Deep learning VADs, trained and optimised for particular degradation conditions, i.e. content

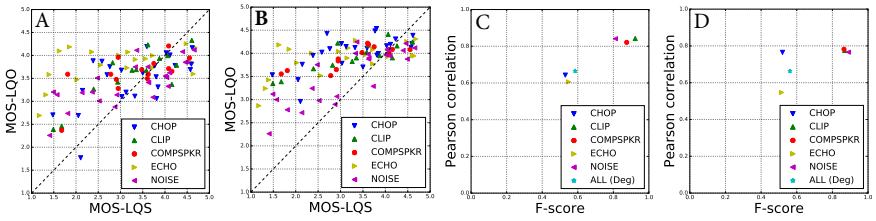

Fig. 4: Subjective vs objective quality prediction (degradations per-condition) of TCD-VoIP (a) No VAD [Baseline]; (b) WS VAD and F-score comparison of best traditional and deep learning VAD with quality correlation for P.563 (Pearson correlation) (c) WS VAD; and (d) bDNN.

sensitive, could have a significant effect on speech quality prediction accuracy as classifying a speech signal by condition and switching VAD to an condition optimal VAD may be easier and more robust than trying to tune traditional signal based VAD threshold parameters. Substituting the two VADs with a single VAD that is aware of content may yield further improvement in speech quality estimation. This provides an opportunity for future work.

6 Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077.

References

1. ITU-T recommendation P.800: Methods for subjective determination of transmission quality (1996)
2. Alcatel-Lucent: PSTN industry analysis and service provider strategies: Synopsis. Alcatel-Lucent, Paris, France, Tech. Rep. Bell Labs Analysis for BT (2013)
3. Bruhn, S., Grancharov, V., Kleijn, W.B.: Low-complexity, non-intrusive speech quality assessment (June 2012), US Patent 8,195,449
4. Garofolo, J.S.: Timit acoustic phonetic continuous speech corpus. Linguistic Data Consortium (1993)
5. Giannakopoulos, T.: A method for silence removal and segmentation of speech signals. University of Athens, Athens **2** (2009)
6. Harte, N., Gillen, E., Hines, A.: TCD-VoIP, a research database of degraded speech for assessing quality in VoIP applications. In: Quality of Multimedia Experience (QoMEX), Seventh International Workshop. pp. 1–6 (2015)
7. Hines, A., Gillen, E., Harte, N.: Measuring and monitoring speech quality for Voice over IP with POLQA, ViSQOL and P.563. INTERSPEECH, Dresden, Germany (2015)
8. Hines, A., Skoglund, J., Kokaram, A., Harte, N.: Monitoring the effects of temporal clipping on VoIP speech quality. In: INTERSPEECH (2013)

9. Hirsch, H.G., Pearce, D.: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW) (2000)
10. Kim, D.S., Tarraf, A.: ANIQUE+: A new American national standard for non-intrusive estimation of narrow-band speech quality. Bell Labs Technical Journal **12**(1), 221–236 (2007)
11. Kim, J., Hahn, M.: Voice activity detection using an adaptive context attention model. IEEE Signal Processing Letters (2018)
12. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arxiv 1412.6980 (2014)
13. Malfait, L., Berger, J., Kastner, M.: P.563-The ITU-T standard for single-ended speech quality assessment. IEEE Transactions on Audio, Speech, and Language Processing **14**(6), 1924–1934 (2006)
14. Möller, S., Chan, W.Y., Côté, N., Falk, T.H., Raake, A., Wältermann, M.: Speech quality estimation: Models and trends. IEEE Signal Processing Magazine **28**(6), 18–28 (2011)
15. Ooster, J., Huber, R., Meyer, B.T.: Prediction of perceived speech quality using deep machine listening. Proc. Interspeech 2018 pp. 976–980 (2018)
16. P.563, I.T.R.: Single-ended method for objective speech quality assessment in narrow-band telephony applications. International Telecommunication Union, Geneva, Switzerland (2004)
17. Pham, T.V., Tang, C.T., Stadtschnitzer, M.: Using artificial neural network for robust voice activity detection under adverse conditions. In: Computing and Communication Technologies. International Conference. pp. 1–8 (2009)
18. Počta, P., Melvin, H., Hines, A.: An analysis of the impact of playout delay adjustments introduced by VoIP jitter buffers on listening speech quality. Acta Acustica united with Acustica **101**(3), 616–631 (2015)
19. Saha, G., Chakroborty, S., Senapati, S.: A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. In: Proceedings of the 11th national conference on communications (NCC). pp. 291–295 (2005)
20. Sak, H., Senior, A., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Fifteenth annual conference of the international speech communication association (2014)
21. Sakhnov, K., Verteletskaya, E., Simak, B.: Approach for energy-based voice detector with adaptive scaling factor. IAENG International Journal of Computer Science **36**(4) (2009)
22. Stork, D.G., Duda, R.O., Hart, P.E., Stork, D.: Pattern classification. A Wiley-Interscience Publication (2001)
23. Vesperini, F., Vecchiotti, P., Principi, E., Squartini, S., Piazza, F.: Deep neural networks for multi-room voice activity detection: Advancements and comparative evaluation. In: Neural Networks (IJCNN), International Joint Conference. pp. 3391–3398 (2016)
24. Zazo Candil, R., Sainath, T.N., Simko, G., Parada, C.: Feature learning with raw-waveform cldnns for voice activity detection (2016)
25. Zhang, X.L., Wang, D.: Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
26. Zhang, X.L., Wang, D.: Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **24**(2), 252–264 (2016)