Hines *et al. Journal on Audio, Speech, and Music Processing* (2015) 2015:13
DOI 10.1186/s13636-015-0054-9

EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

# ViSQOL: an objective speech quality model

Andrew Hines[1,2*], Jan Skoglund[3], Anil C Kokaram[3] and Naomi Harte[2]

**Abstract**

This paper presents an objective speech quality model, ViSQOL, the Virtual Speech Quality Objective Listener. It is a signal-based, full-reference, intrusive metric that models human speech quality perception using a spectro-temporal measure of similarity between a reference and a test speech signal. The metric has been particularly designed to be robust for quality issues associated with Voice over IP (VoIP) transmission. This paper describes the algorithm and compares the quality predictions with the ITU-T standard metrics PESQ and POLQA for common problems in VoIP: clock drift, associated time warping, and playout delays. The results indicate that ViSQOL and POLQA significantly outperform PESQ, with ViSQOL competing well with POLQA. An extensive benchmarking against PESQ, POLQA, and simpler distance metrics using three speech corpora (NOIZEUS and E4 and the ITU-T P.Sup. 23 database) is also presented. These experiments benchmark the performance for a wide range of quality impairments, including VoIP degradations, a variety of background noise types, speech enhancement methods, and SNR levels. The results and subsequent analysis show that both ViSQOL and POLQA have some performance weaknesses and under-predict perceived quality in certain VoIP conditions. Both have a wider application and robustness to conditions than PESQ or more trivial distance metrics. ViSQOL is shown to offer a useful alternative to POLQA in predicting speech quality in VoIP scenarios.

**Keywords:** Objective speech quality; POLQA; P.853; PESQ; ViSQOL; NSIM

## 1 Introduction

Predicting how a user perceives speech quality has become more important as transmission channels for human speech communication have evolved from traditional fixed telephony to Voice over Internet Protocol (VoIP)-based systems. Packet-based networks have compounded the traditional background noise quality issues with the addition of new channel-based degradations. Network monitoring tools can give a good indicator of the quality of service (QoS), but predicting the quality of experience (QoE) for the end user of heterogeneous networked systems is becoming more important as transmission channels for human speech communication have a greater reliance on VoIP. Accurate reproduction of the input waveform is not the ultimate goal, as long as the user perceives the output signal as a high-quality representation of their expectation of the original signal input.

Popular VoIP applications, such as Google Hangouts and Skype, deliver multimedia conferencing over standard computer or mobile devices rather than dedicated video conferencing hardware. End-to-end evaluation of the speech quality delivery has become more complex as the number of variables impacting the signal has expanded. For system development and monitoring purposes, quality needs to be reliably assessed. Subjective testing with human listeners is the ground truth measurement for speech quality but is time consuming and expensive to carry out. Objective measures aim to model this assessment, to give accurate estimates of quality when compared with subjective tests.

PESQ (Perceptual Evaluation of Speech Quality) [1] and the more recent POLQA (Perceptual Objective Listening Quality Assessment) [2], described in ITU standards, are full-reference measures meaning they allow prediction of speech quality by comparing a reference to a received signal. PESQ was developed to give an objective estimate of narrowband speech quality and was later extended to also address wideband speech quality [3]. The newer POLQA model yields quality estimates for narrowband, wideband, and super-wideband speech and

*Correspondence: andrew.hines@dit.ie
[1] School of Computing, Dublin Institute of Technology, Kevin St, Dublin 8, Ireland
[2] Sigmedia, Department of Electronic and Electrical Engineering, Trinity College Dublin, College Green, Dublin 2, Ireland
Full list of author information is available at the end of the article

addresses other limitations in PESQ, specifically time alignment and warped speech. It is slowly gaining more widespread use, so as yet, there has been limited publication of its performance outside of its own development and conformance tests.

This work presents an alternative model, the Virtual Speech Quality Objective Listener, or ViSQOL, which has been developed to be a general full-reference objective speech quality metric with a particular focus on VoIP degradations. The experiments presented compare the performance to PESQ and POLQA and benchmarks their performance over a range of common background noises and warp, clock drift, and jitter VoIP impairments.

The early development of ViSQOL was presented in a paper introducing the model's potential to measure two common VoIP problems: clockdrift and jitter [4]. Further work developed the algorithm and mapped the model output to mean opinion score (MOS) estimates [5]. This work expands on these experiments and presents a detailed description of the algorithm and experimental results for a variety of quality degradations. The model performance is further evaluated against two more simplistic quality metrics as well as the ITU standards PESQ and POLQA.

Section 2 provides a background and sets the context for this research, giving an introduction to subjective and objective speech quality measurement and related research. Sections 3 and 4 introduce and then describe the ViSQOL model architecture. Section 5 describes five experiments, presents details of the tests undertaken and datasets used, and discusses the experimental results. Section 6 summaries the results, and Section 7 concludes the paper and suggests some areas for further model testing and development.

## 2 Background

### 2.1 Speech quality issues with Voice over IP

There are three factors associated with packet networks that have a significant impact on perceived speech quality: delay, jitter (variations in packet arrival times), and packet loss. All three factors stem from the nature of a packet network, which provides no guarantee that a packet of speech data will arrive at the receiving end in time, or even that it will arrive at all [6]. Packet losses can occur both in routers in the network or at the end point when packets arrive too late to be played out. To account for these factors and to ensure a continuous decoding of packets, a jitter buffer is required at the receiving end. The design trade-off for the jitter buffer is to keep the buffering delay as short as possible while minimizing the number of packets that arrive too late to be used. A large jitter buffer causes an increase in the overall delay and decreases the packet loss. A high delay can severely affect the quality and ease of conversation as the wait leads to annoying talker overlap. The ITU-T Recommendation G.114 [7] states that the one-way delay should be kept below 150 ms for acceptable conversation quality. In practice somewhat larger delays can be tolerated, but in general a latency larger than 300 to 400 ms is deemed unacceptable. A smaller buffer decreases the delay but increases the resulting packet loss. When a packet loss occurs, some mechanism for filling in the missing speech must be incorporated. Such solutions are usually referred to as packet loss concealment (PLC) algorithms, see Kim et al. [8] for a more complete review. This can be done by simply inserting zeros, repeating signals, or by some more sophisticated methods utilizing features of the speech signal, e.g., pitch periods. The result of inserting zeros or repeating packets is choppy speech with highly audible discontinuities perceived as clicks. Pitch-based methods instead try finding periodic segments to repeat in a smooth periodic manner during voiced portions of speech. This typically results in high-quality concealment, even though it may sound robotic and buzzy during events of high packet loss. An example of such a pitch-period-based method is the NetEq [9] algorithm in WebRTC, an open-source platform for audio and video communication over the web [10]. NetEq is continuously adapting the playout timescale by adding or reducing pitch periods to not only conceal lost segments but also to reduce built-up delay in the jitter buffer.

Another important aspect which indirectly may affect the quality is clock drift. Whether the communication end-points are gateways or other devices, low-frequency clock drift between the two can cause receiver buffer overflow or underflow. If the clock drift is not detected accurately, delay builds up during a call, so clock drift can have a significant impact on the speech quality. For example, the transmitter might send packets every 20 ms according to its perception of time, while the receiver's perception is that the packets arrive every 20.5 ms. In this case, for every 40th packet, the receiver has to perform a packet loss concealment to avoid buffer underflow. The NetEq algorithm's timescale modification inherently adjusts for clock drift in a continuous sample-by-sample fashion and thereby avoids such step-wise concealment.

### 2.2 Subjective and objective speech quality assessment

Inherently, the judgement of speech quality for human listeners is subjective. The most reliable method for assessment is via subjective testing with a group of listeners. The ITU-T has developed a widely used recommendation (ITU-T Rec. P.800 [11]) defining a procedure for speech quality subjective tests. The recommendation specifies several testing paradigms. The most frequently used is the Absolute Category Rating (ACR) assessment where listeners rate the quality of speech samples into a scale of 1 to 5 (bad, poor, fair, good, and excellent). The ratings for all listeners are averaged to a single score known as a mean opinion score (MOS). With multiple listeners rating a common minimal value of four samples per condition (spoken by two male and two female speakers), subjective testing is time consuming, expensive, and requires strict adherence to the methodology to ensure applicability of results. Subjective testing is impractical for frequent automated software system regression tests or routine network monitoring applications.

As a result, objective test methods have been developed in recent years and remain a topic of active research. This is often seen as surprising considering telephone communications have been around for a century. The advent of VoIP has introduced a range of new technological issues and related speech quality factors that require the adaptation of speech quality models [12]. Objective models are machine executable and require little human involvement for repeatable automated regression tests to be created for VoIP systems. They are useful tools for a wide audience: VoIP application and codec developers can use them to benchmark and assess changes or enhancements to their products; while telecommunications operators can evaluate speech quality throughout their system life cycles from planning and development through to implementation, optimization, monitoring, and maintenance. They are important tools for a range of research disciplines such as human computer interfaces, e.g., speech or speaker recognition, where knowledge of the quality of the test data is important in quantifying their system's robustness to noise [13]. An extensive review of objective speech quality models and their applications can be found in [14].

Objective methods can be classified into two major categories: parameter-based and signal-based methods. Parameter-based methods do not test signals over the channel but instead predict the speech quality through modeling the channel parameters. The E-model is an example of a parameter-based model. It is defined by ITU-T Recommendations G.107 [15] (narrowband version) and G.107.1 [16] (wideband version) and is primarily used for transmission planning purposes in narrowband and wideband telephone networks.

This work concentrates on the other main category, namely signal-based methods. They predict quality based on evaluation of a test speech signal at the output of the channel. They can be divided into two further subcategories, intrusive or non-intrusive. Intrusive signal-based methods use an original reference and a degraded signal, which is the output of the system under test. They identify the audible distortions based on the perceptual domain representation of two signals incorporating human auditory models. Several intrusive models have been developed during recent years. The ITU-T Recommendation P.861 (PSQM), published in 1996, was a first attempt to objectively model human listeners and predict speech quality from subjective listener tests. It was succeeded in 2001 by P.862, commonly known as PESQ, a full-reference metric for predicting speech quality. PESQ has been widely used and was enhanced and extended over the next decade. It was originally designed and tested on narrowband signals. It improved on PSQM and the model handles a range of transmission channel problems and variations including varied speech levels, codecs, delays, packet loss, and environmental noise. However, it has a number of acknowledged shortcomings including listening levels, loudness loss, effects of delay in conversational tests, talker echo, and side tones [1]. An extension to PESQ was developed that adapted the input filters and MOS mapping to allow wideband signal quality prediction [3].

The newer POLQA algorithm, presented in ITU-T P.863 Recommendation, addresses a number of the limitations of PESQ as well as improving the overall correlation with subjective MOS scores. POLQA also implements an 'idealisation' of the reference signal. This means that it will attempt to create a reference signal weighting the perceptually salient data before comparing it to the degraded signal. It allows for predicting overall listening speech quality in two modes: narrowband (300 to 3,400 Hz) and superwideband (50 to 14,000 Hz). It should be noted that in the experiments described in this paper, POLQA was used in narrowband mode where the specification defines the estimated MOS listener quality objective output metric (MOS-LQOn, with *n* signifying narrowband testing) saturating at 4.5.

In contrast to intrusive methods, the idea of the single-ended (non-intrusive) signal-based method is to predict the quality without access to a reference signal. The result of this comparison can further be modified by a parametric degradation analysis and integrated into an assessment of overall quality. The most widely used non-intrusive models include Auditory Non-Intrusive QUality Estimation (ANIQUE+) [17] and ITU-T standard P.563 [18], although it is still an active area of research [19-22].

For much of the published work on speech quality in VoIP, PESQ is used as an objective metrics of speech quality, e.g., [23,24]. PESQ was originally designed with narrowband telephony in mind and did not specifically target the most common quality problems encountered in VoIP systems described in 2.1. POLQA has sought to address some of the known shortcomings of PESQ, but only a small number of recent publications, e.g., [25], have begun to evaluate the performance of POLQA for VoIP issues. PESQ is still worthy of analysis as recently published research continues to use PESQ for VoIP speech quality assessment, e.g., [26,27].

This paper presents the culmination of work from the authors [4,5,28] in developing a new objective metric of speech quality, called ViSQOL. ViSQOL has been designed to be particularly sensitive to VoIP degradation but without sacrificing wider deployability. The metric works by examining similarity in time-frequency representations of the reference and degraded speech, looking for the manifestation of these VoIP events. The new metric is compared to both PESQ and POLQA.

### 2.3 Benchmark models

Both ITU-T models, PESQ and POLQA, involve a complex series of pre-processing steps to achieve a comparison of signals. These deal with factors like loudness levels, temporal alignments, and delays. They also include a perceptual model that filters the signal using bandpass filters to mimic the frequency sensitivity and selectivity of the human ear. For ease of comparison with ViSQOL, block diagrams of the three models are presented in Figures 1, 2, 3, and 4. The models differ in a variety of ways beyond the fundamental distance calculations between signals, including level alignment, voice activity detection, time-alignment, and mapping from an internal metric to a MOS estimate. All three are quite complex in their implementations and more detail on PESQ and POLQA can be found in the relevant ITU-T standards. Further details on ViSQOL follows in Section 3.

When dealing with speech quality degradations that are constrained to background noise or speech enhancement algorithms attempting to counteract noise, simple SNR distance metrics may suffice. This was shown to be the case by Hu and Loizou when evaluating speech enhancement algorithms with a variety of objective quality metrics [29]. However, these metrics have difficulty with modern communications networks. Modern codecs can produce high-quality speech without preserving the input waveform. Quality measures based on waveform similarity do not work for these codecs. Comparing signals in the spectral domain avoids this problem and can produce results that agree with human judgement. The two best performing metrics from Hu and Loizou's study, the log-likelihood ratio (LLR) and frequency domain segmental signal-to-noise ratio (fwSNRSeg) [29,30], are tested along with the specialised speech quality metrics, PESQ and POLQA, to illustrate their strengths and weaknesses.

### 2.4 Experimental datasets

Subjective databases used for metric calibration and testing are a key component in objective model development. Unfortunately, many datasets are not made publicly available; and those that are frequently used do not contain a realistic sample of degradation types targeting a specific application under study, or their limited size does not allow for statistically significant results. MOS scores can vary, based on culture and language, or balance of conditions in a testset, even for tests within the same laboratory [31]. The coverage of the data in terms of variety of conditions and range of perceived quality is usually limited to a range of conditions of interest for a specific research topic. A number of best practice procedures have been set out by the ITU, e.g., the ITU-T P.800 test methodology [11], to ensure statistically reliable results. These cover details such as the number of listeners, environmental conditions, speech sample lengths, and content and help to ensure that MOS scores are gathered and interpreted correctly. This work presents results from tests using a combination of existing databases where available and subjective tests carried out by the authors for assessing objective model performance for a range of VoIP specific and general speech degradations.

## 3 Measuring speech quality through spectrogram similarity

ViSQOL was inspired by prior work on speech intelligibility by two of the authors [32,33]. This work used a model of the auditory periphery [34] to produce auditory nerve discharge outputs by computationally simulating the middle and inner ear. Post-processing of the model outputs yield a neurogram, analogous to a spectrogram with

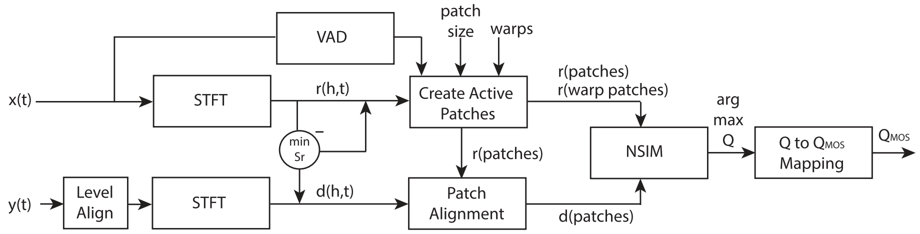

**Figure 1** Block diagram of ViSQOL. High-level block diagram of the ViSQOL algorithm, also summarised in Algorithm 1. Pre-processing includes signal leveling and production of spectrogram representations of the reference and degraded signal. Similarity comparison: alignment, warp compensation, and calculating similarity scores between patches from the spectrograms. Quality prediction: patch similarity scores are combined and translated to an overall objective MOS result. Full reference MATLAB implementation available.

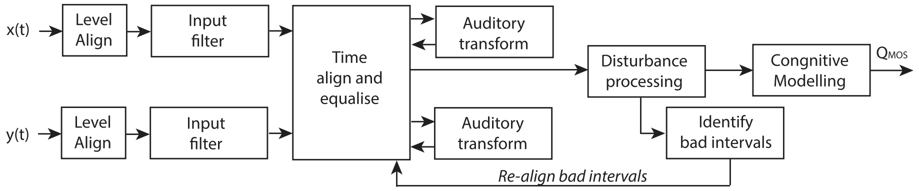

**Figure 2** Block diagram of PESQ. PESQ carries out level alignment, mimics the resolution of the human ear, and carries out alignment to compensate for network delays.

time-frequency color intensity representation related to neural firing activity.

Most speech quality models quantify the degradation in a signal, i.e., the amount of noise or distortion in the speech signal compared to a 'clean' reference. ViSQOL focuses on the similarity between a reference and degraded signal by using a distance metric called the Neurogram Similarity Index Measure or NSIM. NSIM was developed to evaluate the auditory nerve discharges in a full-reference way by comparing the neurogram for reference speech to the neurogram from degraded speech to predict speech intelligibility. It was inspired and adapted for use in the auditory domain from an image processing technique, structural similarity, or SSIM [35], which was created to predict the loss of image quality due to compression artifacts. Adaptations of SSIM have been used to predict audio quality [36] and more recently have been applied in place of simple mean squared error in aeroacoustics [37]. Computation of NSIM is described below in Section 4.2.3.

While speech intelligibility and speech quality are linked, work by Voiers [38] showed that an amplitude-distorted signal that had been peak clipped did not impact intelligibility but seriously affected the quality. This phenomena is well illustrated by examples of vocoded or robotic speech where the intelligibility can be 100% but the quality is ranked as bad or poor. In evaluating the speech intelligibility provided by two hearing aid algorithms with NSIM, it was noted that while the intelligibility level was the same for both, the NSIM predicted higher levels of similarity for one algorithm over the other [39]. This suggested that NSIM may be a good indicator of other factors beyond intelligibility such as speech quality. It was necessary to evaluate intelligibility after the auditory periphery when modeling hearing impaired listeners as the signal impairment occurs in the cochlea. This paper looks at situations where the degradation occurs in the communication channel, and hence assessing the signal directly using NSIM on the signal spectrograms rather than neurograms simplifies the model. This decreased the computational complexity of the model by two magnitudes to an order comparable with other full-reference metrics such as PESQ and POLQA.

## 4 Algorithm description

ViSQOL is a model of human sensitivity to degradations in speech quality. It compares a reference signal with a degraded signal. The output is a prediction of speech quality perceived by an average individual. The model has five major processing stages shown in the block diagram Figure 1: pre-processing; time alignment; predicting warp; similarity comparison; and a post-process mapping similarity to objective quality. The algorithm is also summarized in Algorithm 1. For completeness, the reader should refer to the reference MATLAB source code implementation of the model available for download [40].

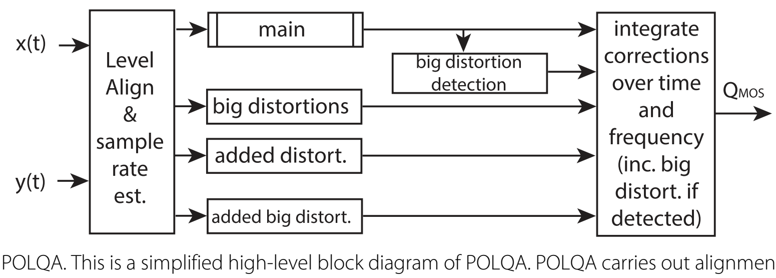

**Figure 3** Block diagram of POLQA. This is a simplified high-level block diagram of POLQA. POLQA carries out alignment per frame and estimates the degraded signal sample rate. The main perceptual model (shown in panel titled 'main' in this figure and detailed in Figure 4) is executed four times with different parameters based on whether big distortions are flagged by the first model. Disturbance densities are calculated for each perceptual model and the integrated model to output a MOS estimate.

**Algorithm 1** Calculate $Q_{MOS} = VISQOL(x, y)$

```
Require: $x$
Require: $y$
Ensure: $dBSPL(y) == dBSPL(x)$
  $r \leftarrow spectrogram(x)$
  $d \leftarrow spectrogram(y)$
  $r \leftarrow r - \arg\min r$
  $d \leftarrow d - \arg\min r$
  for $patch = 1$ to $length(r) - PATCHSIZE$ do
    if $VAD(r(patch)) =$ TRUE then
      $refpatches[\,] \leftarrow r(patch)$
      $refwarppatches[\,] \leftarrow warp(r(patch))$
    end if
    $t_d[\,] \leftarrow alignpatches(refpatches[\,], d)$
  end for
  for all $refpatches$ such that $1 \leq i \leq NUMPATCHES$ do
    for all $warps$ such that $1 \leq w_i \leq NUMWARPS$ do
      for all $t_d$ such that $1 \leq t_i \leq NUMPATCHES$ do
        $q(i) \leftarrow nsim(refpatches(i), d(t_d(t_i))$
        $qwarp(i) \leftarrow nsim(refwarppatches(w_i), d(t_d(t_i))))$

        $q(i) \leftarrow max(q(i), qwarp(i))$
      end for
    end for
  end for
  $Q \leftarrow \sum(q(i))/NUMPATCHES$
  $Q_{MOS} \leftarrow maptomos(Q)$
```

### 4.1 Pre-processing

The pre-processing stage scales the degraded signal $y(t)$, to match the power level of the reference signal $x(t)$. Short-term Fourier transform (STFT) spectrogram representations of the reference and degraded signals are created using critical bands between 150 and 3,400 Hz for narrowband testing and including five further bands to 8,000 Hz for wideband. They are denoted $r$ and $d$, respectively. A 512 sample, 50% overlap periodic Hamming window is used for signals with 16-kHz sampling rate and a 256 sample window for 8-kHz sampling rate to keep frame resolution temporally consistent at 32-ms length with 16-ms spacing. The test spectrograms are floored to the minimum value in the reference spectrogram to level the signals with a 0-dB reference. The spectrograms are used as inputs to the second stage of the model, shown in detail on the right-hand side of Figure 1.

### 4.2 Feature selection and comparison

#### *4.2.1 Time alignment*

The reference signal is segmented into patches for comparison as illustrated in Figure 5. Each patch is 30 frames long (480 ms) by 16 or 21 critical frequency bands [41] (i.e., 150 to 3,400 for narrowband or 50 to 8,000 Hz for wideband signals). A simple energy threshold voice activity detector is used on the reference signal to approximately segment the signal into active patches. NSIM is used to time align the patches to ensure that the patches are aligned correctly even for conditions with high levels of background noise. Each reference patch is aligned with the corresponding area from the test spectrogram. The Neurogram Similarity Index Measure (NSIM) [33] is used to measure the similarity between the reference patch and a test spectrogram patch frame by frame, thus identifying the maximum similarity point for each patch. This is shown in the bottom pane of Figure 5 where each line graphs the NSIM similarity score over time for each patch in the reference signal compared with the example signal. The NSIM at the maxima are averaged over the patches to yield the metric for the example signal.

#### *4.2.2 Predicting warp*

NSIM is more sensitive to time warping than a human listener. The ViSQOL model exploits this by warping the spectrogram patches temporally. It creates alternative reference patches 1% and 5% longer and shorter than the

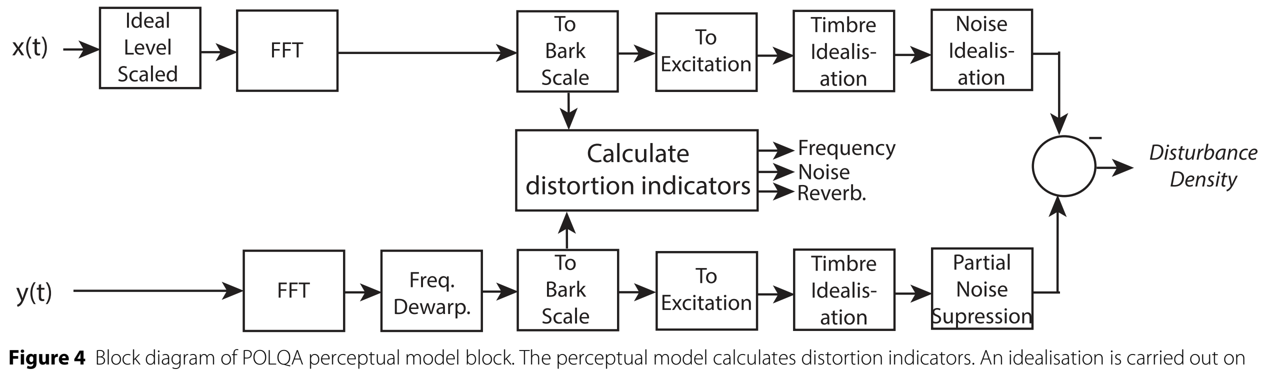

**Figure 4** Block diagram of POLQA perceptual model block. The perceptual model calculates distortion indicators. An idealisation is carried out on the reference signal to remove low levels of noise and optimize timbre of the reference signal prior to the difference calculation for disturbance density estimation.

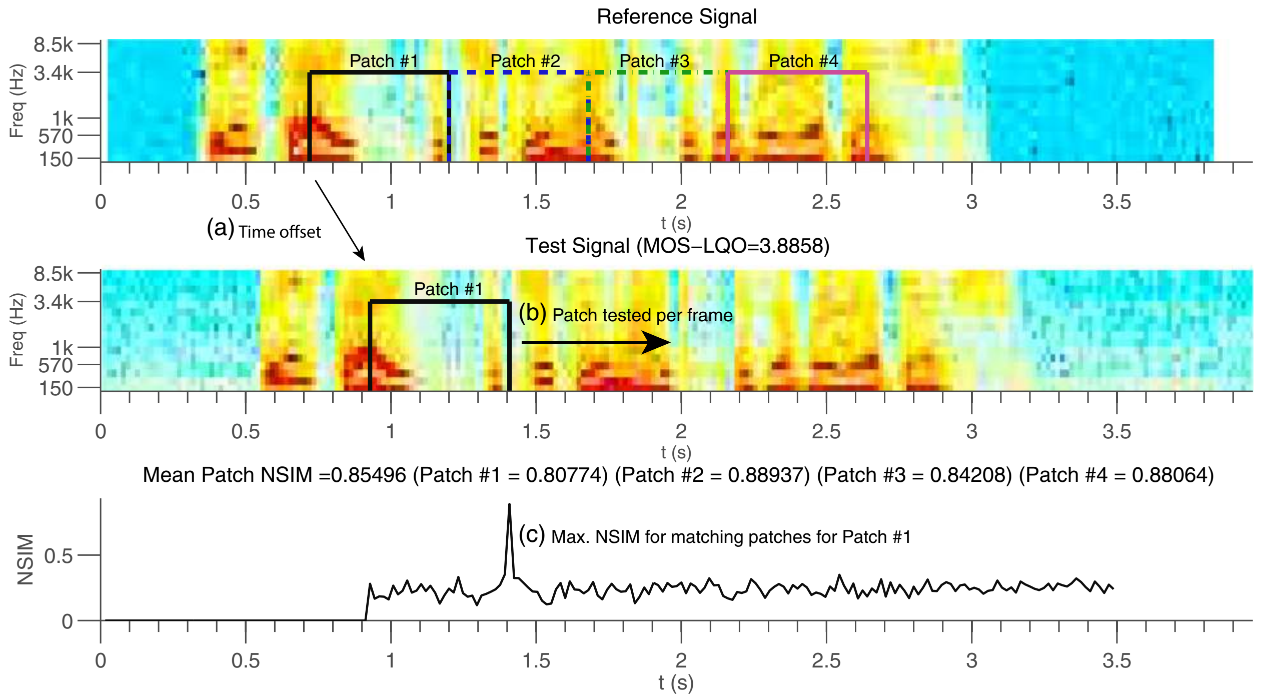

**Figure 5** Speech signals with sample patches. The bottom plot shows the NSIM similarity score for each patch from the reference compared frame by frame across the degraded signal. The NSIM score is the mean of the individual patch scores given in parenthesis. **(a)** Time offset between reference and test signal. **(b)** Patch tested per frame. **(c)** Maximum NSIM for matching patches for Patch # 1.

original reference. The patches are created using a cubic two-dimensional interpolation. The comparison stage is completed by comparing the test patches to the reference patches and all of the warped reference patches using NSIM. If a warped version of a patch has a higher similarity score, this score is used for the patch. This is illustrated in Figure 6.

### *4.2.3 Similarity comparison*

In this work, spectrograms are treated as images to compare similarity. Prior work [32,33] demonstrated that the structural similarity index (SSIM) [35] could be used to discriminate between reference and degraded images of speech to predict intelligibility. SSIM was developed to evaluate JPEG compression techniques by assessing image similarity relative to a reference uncompressed image. It exhibited better discrimination than basic point-to-point measures, i.e., relative mean squared error (RMSE). SSIM uses the overall range of pixel intensity for the image along with a measure of three factors on each individual pixel comparison. The factors, luminance, contrast, and structure, give a weighted adjustment to the similarity measure that looks at the intensity (luminance), variance (contrast), and cross-correlation (structure) between a given pixel and those that surround it versus the reference image. SSIM between two spectrograms, the reference, *r*, and the degraded, *d*, is defined with a weighted function of intensity, *l*, contrast, *c*, and structure, *s*, as

$$S(r,d) = l(r,d)^{\alpha} \cdot c(r,d)^{\beta} \cdot s(r,d)^{\gamma} \qquad (1)$$

$$S(r,d) = \left(\frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1}\right)^{\alpha} \cdot \left(\frac{2\sigma_r\sigma_d + C_2}{\sigma_r^2 + \sigma_d^2 + C_2}\right)^{\beta} \times \left(\frac{\sigma_{rd} + C_3}{\sigma_r\sigma_d + C_3}\right)^{\gamma} \qquad (2)$$

Components are weighted with $\alpha$, $\beta$, and $\gamma$ where all are set to 1 for the basic version of SSIM. Intensity looks at a comparison of the mean, $\mu$, values across the two spectrograms. The structure uses the standard deviation, $\sigma$, and is equivalent to the correlation coefficient between the two spectrograms. In discrete form, $\sigma_{rd}$ can be estimated as

$$\sigma_{rd} = \frac{1}{N-1}\sum_{i=1}^{N}(r_i - \mu_r)(d_i - \mu_d). \qquad (3)$$

where *r* and *d* are time-frequency matrices summed across both dimensions. Full details of calculating SSIM are presented in [35].

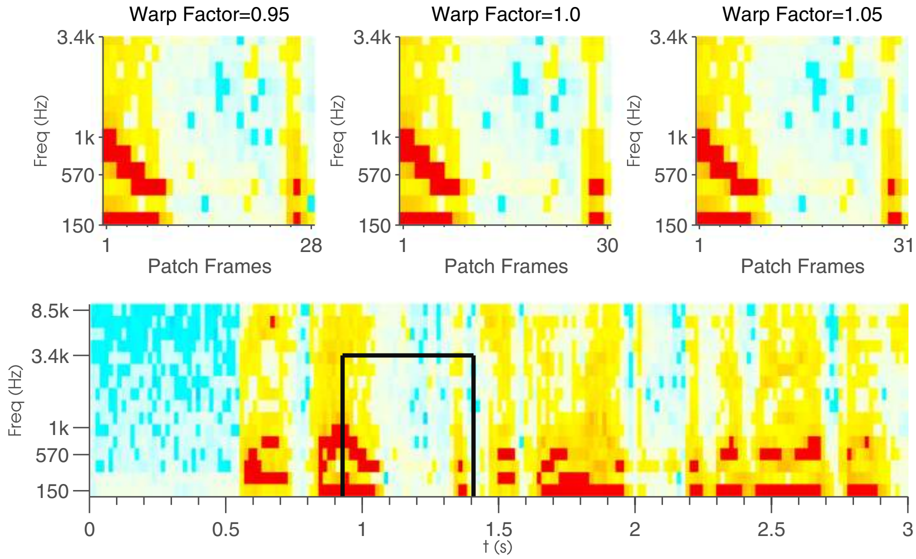

**Figure 6** Patch warping. The versions of the reference patch #1 are shown: warped temporally to 0.95 times the length, un-warped (1.0 factor) and 1.05 times warped. These are compared to the degraded signal at the area of maximum similarity and adjacent frames. The highest similarity score for all warps tested is used for each given patch.

The Neurogram Similarity Index Measure (NSIM) is a simplified version of SSIM that has been shown to perform better for speech signal comparison [33] and is defined as

$$Q(r,d) = l(r,d) \cdot s(r,d) = \frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1} \cdot \frac{\sigma_{rd} + C_3}{\sigma_r.\sigma_d + C_3} \quad (4)$$

As with SSIM, each component also contains constant values $C_1 = 0.01L$ and $C_2 = C_3 = (0.03L)^2$, where $L$ is the intensity range (as per [35]) of the reference spectrogram, which have negligible influence on the results but are used to avoid instabilities at boundary conditions, specifically where $\mu_r^2 + \mu_d^2$ is very close to zero. It was previously established that for the purposes of neurogram comparisons for speech intelligibility estimation, the optimal window size was a 3 × 3 pixel square covering three frequency bands and a 12.8-ms time window [32]. SSIM was further tuned, and it was established that the contrast component provided negligible value when comparing neurograms and that closer fitting to listener test data occurred using only a luminance and structural comparison [33]. Strictly, NSIM has a bounded range $-1 \leq Q \leq 1$ but for spectrograms where the reference is clean speech, the range can be considered to be $0 \leq Q \leq 1$. Comparing a signal with itself will yield an NSIM score of 1. When calculating the overall similarity, the mean NSIM score for the test patches is returned as the signal similarity estimate.

### 4.3 Mapping similarity to objective quality

A mapping function, roughly sigmoid in nature, is used to translate the NSIM similarity score into a MOS-LQOn score and mapped in the range 1 to 5. The mean of the third-order polynomial fitting functions for three of the ITU-T P. Supplemental 23 databases was used to create the mapping function. The database contains test results from a number of research laboratories. Results from three laboratories were used to train the mapping function (specifically those labeled A, C, and D), and laboratory O results were kept aside for metric testing and evaluation. The transfer function, $Q_{MOS} = f(z)$, where $z$ maps the NSIM score, $Q$, to $Q_{MOS}$ is described by

$$\text{clamp}(Q_{\text{MOS}}, a, b) = \begin{cases} m & \text{if } f(z) \leq m, \\ f(z) & a < f(z) \leq n \\ n & \text{if } f(z) > n \end{cases} \quad (5)$$

where $Q_{\text{MOS}} = az^3 + bz^2 + cz + d$, $m = 1$, $n = 5$ and the coefficients are $a = 158.7$, $b = -373.6$, $c = 295.5$ and $d = -75.3$. This transfer function is used for all data tested. A further linear regression fit was applied to the results from all of the objective metrics tested to map the objective scores to the subjective test databases used for evaluation. The correlation statistics are quoted with and without this regression fit.

### 4.4 Changes from early model design

An earlier prototype of the ViSQOL model was presented in prior work [4]. A number of improvements were subsequently applied to the model. Firstly, an investigation of cases with mis-aligned patches was undertaken. While NSIM is computationally more intensive than other alignment techniques such as relative mean squared error (used in [4]), it was found to be more robust [5]. Further experimentation found that while this was sufficient in medium SNR scenarios, RMSE was not robust to SNR levels less than 5 dB and resulted in mis-alignments. An example is presented in Figure 7 where a reference patch containing the utterance ‘days’ is shown along with the same patch from three degraded versions for the same speech sample. The RMSE remains constant for all three while the NSIM score drops in line with the perceptual MOS scores. Secondly, the warping of patches was limited to a 1% and 5% warp compared with earlier tests [4]. This was done for efficiency purposes and did not reduce accuracy.

An efficiency optimization used in the early prototype was found to reduce the accuracy of the prototype and was removed. This change was prompted by poor estimation of packet loss conditions with the earlier model for the dataset used in Experiment 4 below and is a design change to the model rather than training with a particular dataset. Specifically, the earlier model based the quality estimation on the comparison of three patches selected from the reference signal regardless of signal duration. Removing this limitation and using a voice activity detector on the reference signal ensured that all active areas of speech are evaluated. This change ensured that temporally occurring degradations such as packet loss are captured by the model.

Finally, the intensity range, $L$, used by Equation 4 was set locally per patch for the results published in [5]. This was found to offset the range of the quality prediction due to dominance of the $C_1$ and $C_3$ constants in 4. By setting $L$ globally to the intensity range of the reference spectrogram rather than each individual patch, the robustness of NSIM to MOS-LQO mapping across datasets was improved.

## 5 Performance evaluation

The effectiveness of the ViSQOL model is demonstrated with performance evaluation with five experiments covering both VoIP specific degradations and general quality issues. Experiment 1 expands on the results on clock drift and warp detection presented in [5] and includes a comparison with subjective listener data. Experiment 2 evaluates the impact of small playout adjustments due to jitter buffers on objective quality assessment. Experiment

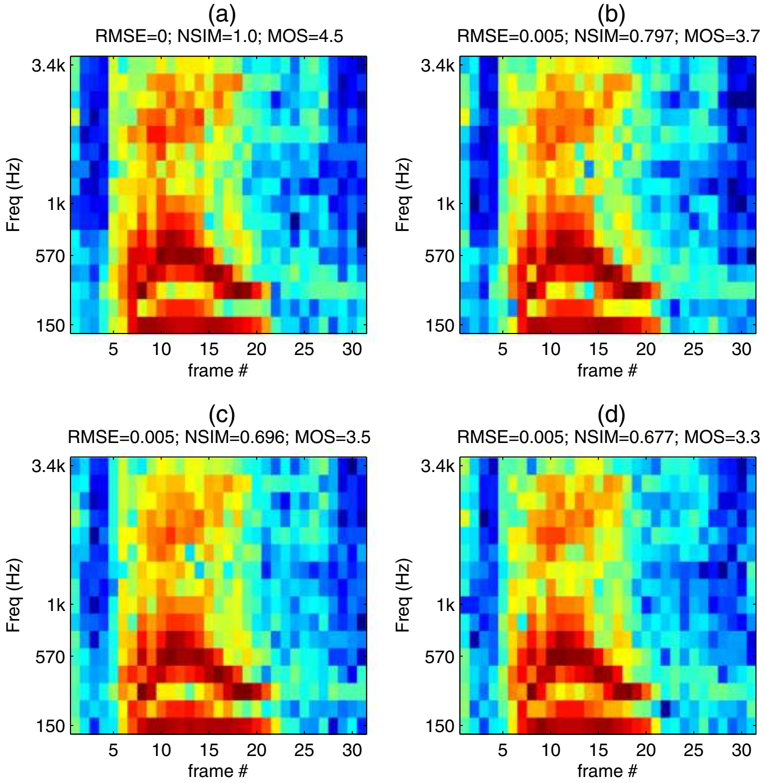

**Figure 7** NSIM and RMSE comparison. **(a)** Reference signal and three progressively degraded signals **(b)** to **(d)**. RMSE scores all degraded signals equally while NSIM shows them to be progressively worse, as per the MOS results.

3 builds upon this to further analyze an open question from [28,42], where POLQA and ViSQOL show inconsistent quality estimations for some combinations of speaker and playout adjustments. Experiment 4 uses a subjectively labeled database of VoIP degradations to benchmark model performance for clock drift, packet loss, and jitter. Finally, Experiment 5 presents benchmark tests with other publicly available speech quality databases to evaluate the effectiveness of the model to a wider range of speech quality issues.

### 5.1 Experiment 1: clock drift and temporal warping

The first experiment tested the robustness of the three models to time warping. Packet loss concealment algorithms can effectively mask packet loss by warping speech samples with small playout adjustments. Here, ten sentences from the IEEE Harvard Speech Corpus were used as reference speech signals [43]. Time warp distortions of signals due to low-frequency clock drift between the signal transmitter and receiver were simulated. The 8-kHz sampled reference signals were resampled to create time-warped versions for resampling factors ranging from 0.85 to 1.15. This test corpus was created specifically for these tests, and a subjective listener test was carried out using ten subjects (seven males and three females) in a quiet environment using headphones. They were presented with 40 warped speech samples and asked to rate them on a MOS ACR scale. The test comprised four versions each of the ten sentences and there were ten resampling factors tested, including a non-resampled factor of 1.

The reference and resampled degraded signal were evaluated using PESQ, POLQA, and ViSQOL for each sentence at each resampling factor. The results are presented in Figure 8. They show the subjective listener test results in the top plot and predictions from the objective measures below. The resample factors from 0.85 to 1.15 along the x-axis are plotted against narrowband mean opinion scores (MOS-LQSn) for the subjective tests and narrowband objective mean opinion scores (MOS-LQOn) quality predictions for the three metrics.

The number of subjects and range of test material in the subjective tests (40 samples with ten listeners) make detailed analysis of the impact of warp on subjective speech quality unfeasible. However, the strong trend visible does allow comparison and comment on the predictive capabilities of the objective metrics.

The subjective results show a large perceived drop off in speech quality for warps of 10% to 15%, but the warps less than 5% seem to suggest a perceptible change but not a large drop in MOS-LQSn score. There is an apparent trend indicating that warp factors less than 1 yield a better quality score than those greater than 1 but further experiments with a range of speakers would be required to rule out voice variability.

The most notable results can been highlighted by examining the plus and minus 5%, 10%, and 15% warp factors. At 5%, the subjective tests point towards a perceptible change in quality, but one that does not alter the MOS-LQSn score to a large extent. ViSQOL predicts a slow drop in quality between 1% and 5%, and POLQA predicts no drop. Either result would be preferred to those of PESQ which predicts a rapid drop to just above 1 MOS-LQOn for a warp of 5%.

At 10% to 15%, the subjective tests indicate that a MOS-LQSn of 2 to 3 should be expected and ViSQOL predicts this trend. However, both POLQA and PESQ have saturated their scale and predict a minimum MOS-LQOn score of 1% from 10% warping. Warping of this scale does cause a noticeable change in the voice pitch from the reference speech but the gentle decline in quality scores predicted by ViSQOL is more in line with listeners' opinions than those of PESQ and POLQA.

The use of jitter buffers is ubiquitous in VoIP systems and often introduces warping to speech. The use of NSIM for patch alignment combined with estimating the similarity using warp-adjusted patches provides ViSQOL with a promising warp estimation strategy for speech quality estimation. Small amounts of warp (around 5% or less) are critical for VoIP scenarios, where playout adjustments are commonly employed. Unlike PESQ where small warps cause large drops in predicted quality, both POLQA and ViSQOL exhibit a lack of sensitivity for warps up to 5% that reflect the listener quality experience.

### 5.2 Experiment 2: playout delay changes

Short network delays are commonly dealt with using per talkspurt adjustments, i.e., inserting or removing portions of silence periods, to cope with time alignment in VoIP. Work by Pocta et al. [42] used sentences from the English speaking portion of ITU-T P Supplement 23 coded-speech database [44] to develop a test corpus of realistic delay adjustment conditions. One hundred samples (96 degraded and four references, two male and two female speakers) covered a range of 12 realistic delay adjustment conditions. The adjustments were a mix of positive and negative adjustments summing to zero (adding and removing silence periods). The conditions comprised two variants (A and B) with the adjustments applied towards the beginning or end of the speech sample. The absolute sum of adjustments ranged from 0 to 66 ms. Thirty listeners participated in the subjective tests, and MOS scores were averaged for each condition.

Where Experiment 1 investigated time warping, this experiment investigates a second VoIP factor, playout delay adjustments. They are investigated and presented here as isolated factors rather than combined in a single test. In a real VoIP system, the components would occur

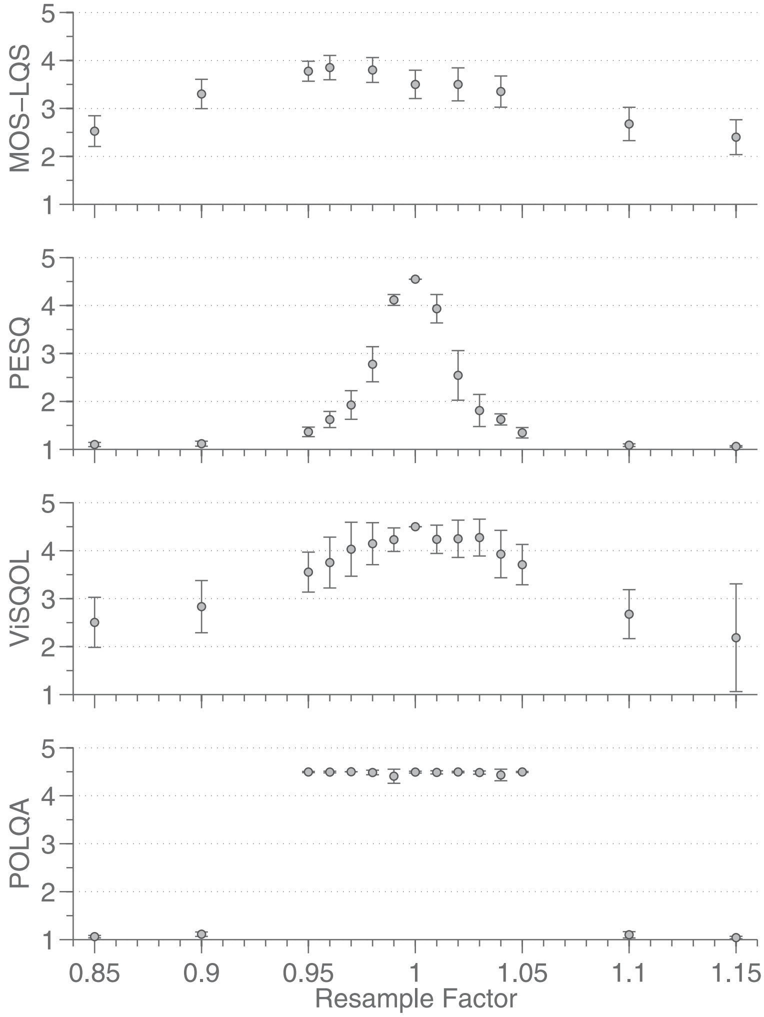

**Figure 8** Experiment 1: clock drift and warp test. Subjective MOS-LQS results for listener tests with MOS-LQOn predictions below for each model comparing ten sentences for each resample factor.

together but as a practical compromise, the analysis is performed in isolation.

The adjustments used are typical (in extent and magnitude) of those introduced by VoIP jitter buffer algorithms [45]. The subjective test results showed that speaker voice preference dominated the subjective test results more than playout delay adjustment duration or location [42]. By design, full-reference objective metrics, including ViSQOL, do not qualify speaker voice difference reducing their correlation with the subjective tests.

The test conditions were compared to the reference samples for the 12 conditions, and the results for ViSQOL, PESQ, and POLQA were compared to those from the subjective tests. These tests and the dominant subjective factors are discussed in more detail in [28,42].

This database is examined here to investigate whether realistic playout adjustments that were shown to be imperceptible from a speech quality perspective are correctly disregarded by ViSQOL, PESQ, and POLQA.

The per condition results previously reported [42] showed that there was poor correlation between subjective and objective scores for all metrics tested but this was as a result of the playout delay changes not being a dominant factor in the speech quality. The results were analyzed for PESQ and POLQA [42] and subsequently for ViSQOL [28], showing MOS scores grouped by speaker and variant instead of playout condition. The combined results from both studies are presented in Figure 9. Looking at the plot of listener test results, the MOS-LQS is plotted on the y-axis against the speaker/variant on the x-axis. It is apparent from the 95% confidence interval bars that condition variability was minimal, and that there was little difference between variants. The dominant factor was the voice quality, i.e., the inherent quality

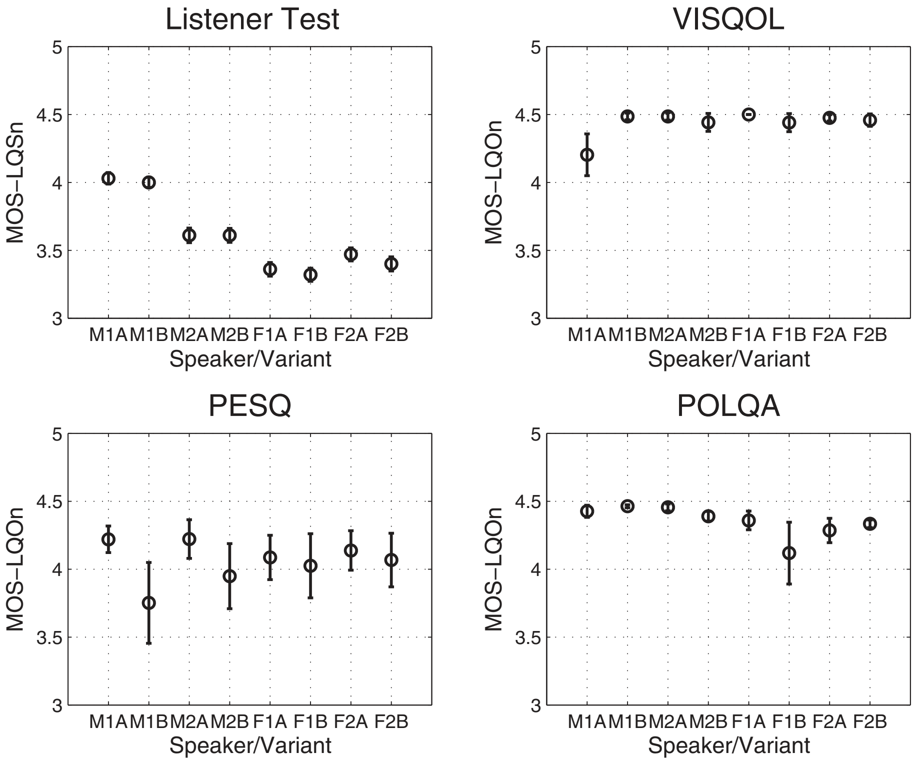

**Figure 9** Experiment 2: playout adjustments. MOS-LQOn predictions for each model broken down by Speaker and delay location variant.

pleasantness of the talker's voice, and not related to transmission factors. Hence, as voice quality is not accounted for by the full-reference metrics, maximum scores should be expected for all speakers. PESQ exhibited variability across all tests, indicating that playout delay was impacting the quality predictions. This was clearly shown in [42]. The results for ViSQOL and POLQA are much more promising apart from some noticeable deviations e.g., the Male 1, Variant A (M1A) for ViSQOL; and the Female 1, Variant B (F1B) for POLQA.

### 5.3 Experiment 3: playout delay changes II

A follow-up test was carried out to try and establish the cause of the variability in results from Experiment 2. This test focused on two speech samples from Experiment 2 where ViSQOL and POLQA predicted quality to be much lower than was found with subjective testing.

For this experiment, two samples were examined. In the first, a silent playout adjustment is inserted in a silence period and in the second, it is inserted within an active speech segment. The start times for the adjustments are illustrated in the lower panes of Figure 10. The quality was measured for each test sentence containing progressively longer delay adjustments. The delay was increased from 0 to 40 ms in 2-ms increments. The upper panes present the results with the duration of the inserted playout adjustment on the x-axis against the predicted MOS-LQOn from POLQA and ViSQOL on the y-axis.

ViSQOL displays a periodic variation of up to 0.5 MOS for certain adjustment lengths. Conversely, POLQA remains consistent in the second test (aside from a small drop of around 0.1 for a 40-ms delay), while in the first test, delays from 4 up to 14 ms cause a rapid drop in predicted MOS with a maximum drop in MOS-LQOn of almost 2.5. These tests highlight the fact that not all imperceptible signal adjustments are handled correctly by either model.

The ViSQOL error is down to the spectrogram windowing and the correct alignment of patches. The problems highlighted by the examples shown here occur only in specific circumstances where the delays are of certain lengths. Also, as demonstrated by the results in the previous experiment, the problem can be alleviated by a canceling effect of multiple delay adjustments where positive and negative adjustments balance out the mis-alignment.

Combined with warping, playout delay adjustments are a key feature for VoIP quality assessment. Flagging these two imperceptible temporal adjustments as a quality issue could mask other factors that actually are perceptible. Although both have limitations, ViSQOL and POLQA are again performing better than PESQ for these conditions.

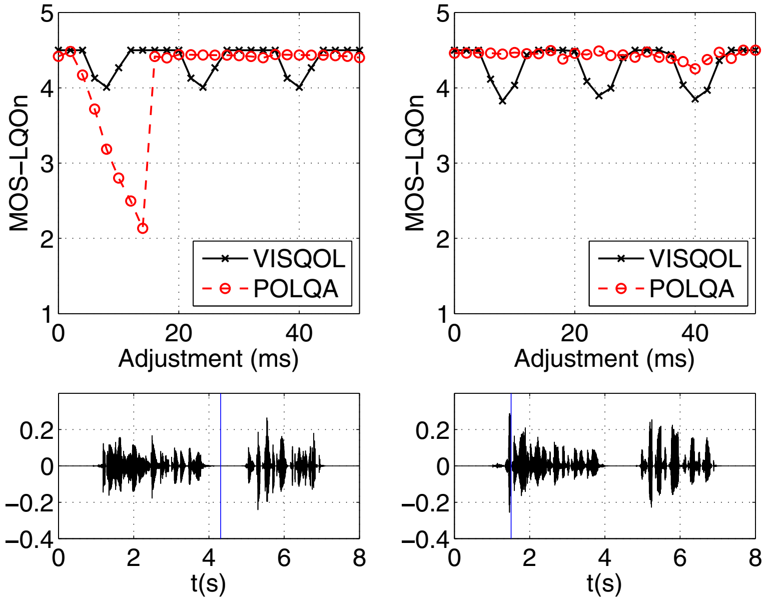

**Figure 10** Experiment 3: progressive playout delays. Above, objective quality predictions for progressively increasing playout delays using two sample sentences. Below, sample signals with playout delay locations marked.

### 5.4 Experiment 4: VoIP specific quality test

A VoIP speech quality corpus, referred to in this paper as the GIPS E4 corpus, contains tests of the wideband codec iSAC [46] with superwideband references. The test was a MOS ACR listening assessment, performed in Native British English. Within these experiments, the iSAC wideband codec was assessed with respect to speech codec and condition. The processed sentence pairs were each scored by 25 listeners. The sentences are from ITU-T Recommendation P.501 [47] which contains two male and two female (British) English speakers sampled at 32 kHz.

For these tests, all signals were down-sampled to 8-kHz narrowband signals. Twenty-seven conditions from the corpus were tested with four speakers per condition (two males and two females). Twenty-five listeners scored each test sample, resulting in 100 votes per condition. The breakdown of conditions was as follows: 10 jitter conditions, 13 packet losses, and four clock drifts. The conditions cover real time, 20 kbps and 32 kbps versions of the iSAC codec. Details of the conditions in the E4 database are summarized in Table 1. While the corpus supplied test files containing the four speakers' sentences concatenated together for each condition, they were separated and tested individually with the objective measures. This dataset contains examples of some of the key VoIP quality degradations that ViSQOL was designed to accurately estimate as jitter, clock drift, and packet loss cause problems with time-alignment and signal warping that are specifically handed by the model design.

The results are presented in Figure 11. The scatter of conditions highlights that PESQ tended to under-predict and POLQA tended to over-predict the MOS scores for the conditions while the ViSQOL estimates were more tightly clustered. Correlation scores for all metrics are presented in Table 2.

### 5.5 Experiment 5: non-VoIP specific quality tests

A final experiment used two publicly available databases to give an indication of ViSQOL's more general speech quality prediction capabilities.

The ITU-T P Supplement 23 (P.Sup23) coded-speech database was developed for the ITU-T 8 kbit/s codec (Recommendation G.729) characterization tests [44]. The conditions are exclusively narrowband speech degradations but are useful for speech quality benchmarking and remain actively used for objective VoIP speech quality models, e.g., [48]. It contains three experimental datasets with subjective results from tests carried out in four labs. Experiment 3 in [44] contains four speakers (two males and two females) for 50 conditions covering a range of VoIP degradations and was evaluated using ACR. The reference and degraded PCM speech material and subjective scores are provided with the database. The English language data (lab O) is referred to in this paper as the P.Sup23 database. As stated in Section 4.3, the subjective results from the other labs (i.e., A, B, and D) were used in the model design for the similarity score to objective quality mapping function.

NOIZEUS [49] is a narrowband 8-kHz sampled noisy speech corpus that was originally developed for evaluation

**Table 1 GIPS E4 database**

| Cond no. | Bitrate | Condition | Cond no. | Bitrate | Condition | Cond no. | Bitrate | Condition |
|---|---|---|---|---|---|---|---|---|
| 1 | Real-time | Jitter | 10 | 20 kbps | Jitter | 19 | 32 kbps | Clock drift |
| 2 | Real-time | Jitter | 11 | 20 kbps | Jitter | 20 | 32 kbps | Jitter |
| 3 | Real-time | Packet loss | 12 | 20 kbps | Jitter | 21 | 32 kbps | Jitter |
| 4 | Real-time | Packet loss | 13 | 20 kbps | Jitter | 22 | 32 kbps | Jitter |
| 5 | Real-time | Packet loss | 14 | 20 kbps | Packet loss | 23 | 32 kbps | Jitter |
| 6 | Real-time | Packet loss | 15 | 20 kbps | Packet loss | 24 | 32 kbps | Packet loss |
| 7 | Real-time | Packet loss | 16 | 20 kbps | Packet loss | 25 | 32 kbps | Packet loss |
| 8 | 20 kbps | Clock drift | 17 | 32 kbps | Packet loss | 26 | 32 kbps | Packet loss |
| 9 | 20 kbps | Clock drift | 18 | 32 kbps | Clock drift | 27 | 32 kbps | Packet loss |

Tests conditions and bitrates using iSAC codec.

of speech enhancement algorithms. Mean opinion scores (MOSs) for a subset of the corpus were obtained using the ITU-T Recommendation P.835 [50] methodology for subjective evaluation. It uses three ratings for each speech sample: the quality of the speech signal alone on a 5-point scale; the intrusiveness of the background noise on a 5-point scale; and the overall signal quality as a MOS ACR. This method was designed to reduce a listener's uncertainty as to the source of the quality issue, e.g., is it the speech signal itself that has been muffled or otherwise impaired or is it a background noise or a combination of both. Further work carried out by Hu and Loizou studied the correlation between objective measures and the subjective quality of noise-suppressed speech [29] and compared PESQ with a range of segmental SNR, LPC, and distance metrics. For the experiments in this paper, only the overall MOS scores were analyzed. Speech subjected to enhancement algorithms, as in the NOIZEUS database, was omitted from the validated scope of POLQA and PESQ. Although the NOIZEUS dataset was not included in the validation testing of POLQA, the specification does not specifically exclude voice enhancement, as was the case for PESQ [25].

Four noise types from the full NOIZEUS corpus were tested: babble, car, street, and train. Each noise type was tested with 13 speech enhancement algorithms plus the

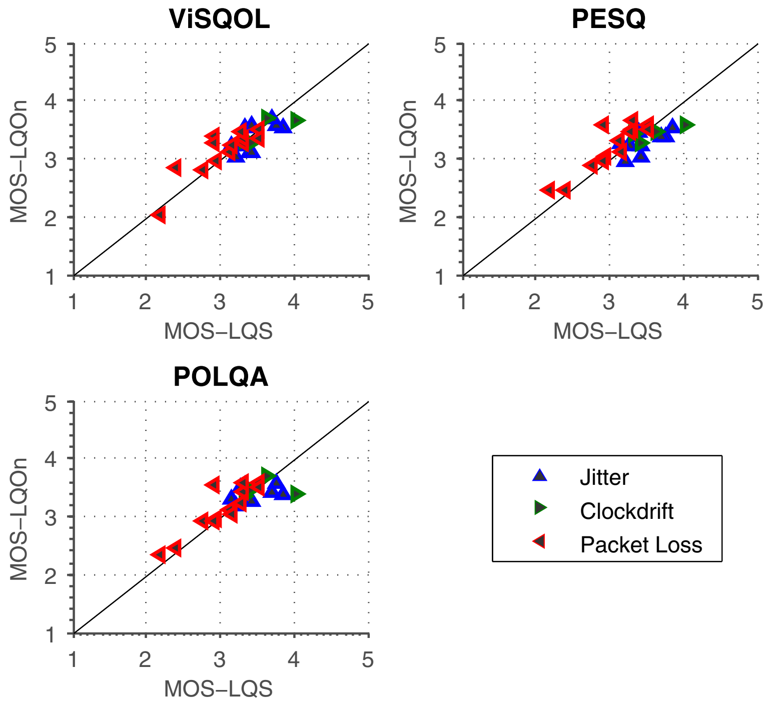

**Figure 11** Experiment 4: scatter results for GIPS E4 VoIP database. Objective measures against subjective MOS scores for VoIP degradations. Plotted after linear regression fitting.

**Table 2 Statistics for Experiments 4 and 5**

| | E4 | | | NOIZEUS | | | P.Sup23 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Pearson** | **Spearman** | **Std. err.** | **Pearson** | **Spearman** | **Std. err.** | **Pearson** | **Spearman** | **Std. err.** |
| Without fit | | | | | | | | | |
| ViSQOL | 0.80 | 0.74 | 0.20 | 0.87 | 0.74 | 0.23 | 0.77 | 0.64 | 0.47 |
| PESQ | 0.78 | 0.57 | 0.18 | 0.90 | 0.57 | 0.20 | 0.92 | 0.90 | 0.29 |
| POLQA | 0.81 | 0.65 | 0.26 | 0.77 | 0.65 | 0.29 | 0.96 | 0.96 | 0.20 |
| LLR | 0.25 | 0.27 | 0.20 | 0.88 | 0.27 | 0.22 | 0.44 | 0.18 | 0.65 |
| fwSNRSeg | 0.21 | 0.14 | 0.21 | 0.86 | 0.14 | 0.24 | 0.48 | 0.38 | 0.64 |
| With linear fit | | | | | | | | | |
| ViSQOL | 0.85 | 0.77 | 0.22 | 0.84 | 0.86 | 0.25 | 0.81 | 0.70 | 0.43 |
| PESQ | 0.78 | 0.57 | 0.26 | 0.90 | 0.88 | 0.20 | 0.92 | 0.90 | 0.29 |
| POLQA | 0.81 | 0.65 | 0.24 | 0.77 | 0.79 | 0.29 | 0.96 | 0.96 | 0.20 |
| LLR | 0.25 | 0.27 | 0.40 | 0.88 | 0.88 | 0.22 | 0.44 | 0.18 | 0.65 |
| fwSNRSeg | 0.21 | 0.14 | 0.40 | 0.86 | 0.85 | 0.24 | 0.48 | 0.38 | 0.64 |

noisy non-enhanced speech at two SNR levels (5 and 10 dB). This gave a total of 112 conditions (four noise types, 14 enhancement variations and two SNR levels). Thirty-two listeners rated the overall quality for each condition with 16 sentences. The MOS scores were averaged for listeners and sentences across each condition. For objective metric testing, the results were calculated in a corresponding manner, with a mean score for the 16 sentences calculated per condition.

Hu and Loizou [29] used the NOIZEUS database to evaluate seven objective speech quality measures. They also investigated composite measures by combining other measures in a weighted manner with PESQ as they did not expect simple objective measures to correlate highly with signal/noise distortion *and* overall quality. The methodology in this work follows the same experiment design and performance evaluation as Hu and Loizou [29]. They measured Pearson's correlation coefficient across the 112 conditions for each measure as well as the standard deviation of the error. For predicting overall quality, they found that PESQ generated the highest correlation of the metrics tested. Absolute values of Pearson's correlation coefficient, $|\rho|$, can be calculated using

$$\rho = \frac{\sum_i (o_i - \bar{o})(s_i - \bar{s})}{\sqrt{\sum_i (o_i - \bar{o})^2}\sqrt{\sum_i (s_i - \bar{s})^2}} \tag{6}$$

where $i$ is the condition index, $o$ is the objective metric score, $s$ is the subjective quality rating (MOS) score, and $\bar{o}$ and $\bar{s}$ are the mean values of $o$ and $s$, respectively. The standard deviation of the error, $\hat{\sigma}_e$, was also measured as a secondary test,

$$\hat{\sigma}_e = \hat{\sigma}_s \sqrt{1 - \rho^2} \tag{7}$$

where $\hat{\sigma}_s$ is the standard deviation of the subjective quality scores, and $s$ and $\rho$ is the correlation coefficient. The Spearman rank correlation was also computed, replacing the quality scores $o$ and $s$ in 6 with their ranks. Hu and Loizou [29] split their data for training and testing. Subsequent evaluations by Kressner et al. [51] repeated the experiments using the full dataset of 1,792 speech files, which is the approach adopted in this study.

The NOIZEUS and P.Sup23 corpora were tested with ViSQOL, PESQ, POLQA, and two additional simple objective metrics, LLR and fwSNRSeg (details of which can be found in [29]). Results were averaged by condition and compared to the average MOS scores per condition. Figure 12 shows the results for each objective quality measure. The scatter shows 112 NOIZEUS conditions and 50 P.Sup23 conditions. The statistical analysis is summarised in Table 2.

As noted by Hu and Loizou in their tests [29], the two less complex metrics, LLR and fwSNRSeg, performed almost as well as PESQ in estimating the quality for the range of background noises evaluated. While they exhibit good correlation for the NOIZEUS tests, their correlation with MOS quality scores for the P.Sup23 and E4 database is much lower (see Table 2). As these are simple measures, it is understandable that while they may perform well for background noise, even if it is not homogeneous, they perform poorly when quantifying more subtle and temporally short-quality degradations such as packet loss or jitter. LLR and fwSNRSeg are simple distance metrics and do not perform any signal alignment, only signal comparison. They have no temporal alignment of signals, leveling, or other pre-processing steps before comparison. They were included in this test to highlight their limitations for VoIP speech quality conditions, and the lack of correlation in the Figure 12 scatter plots illustrates the performance variability between the difference datasets.

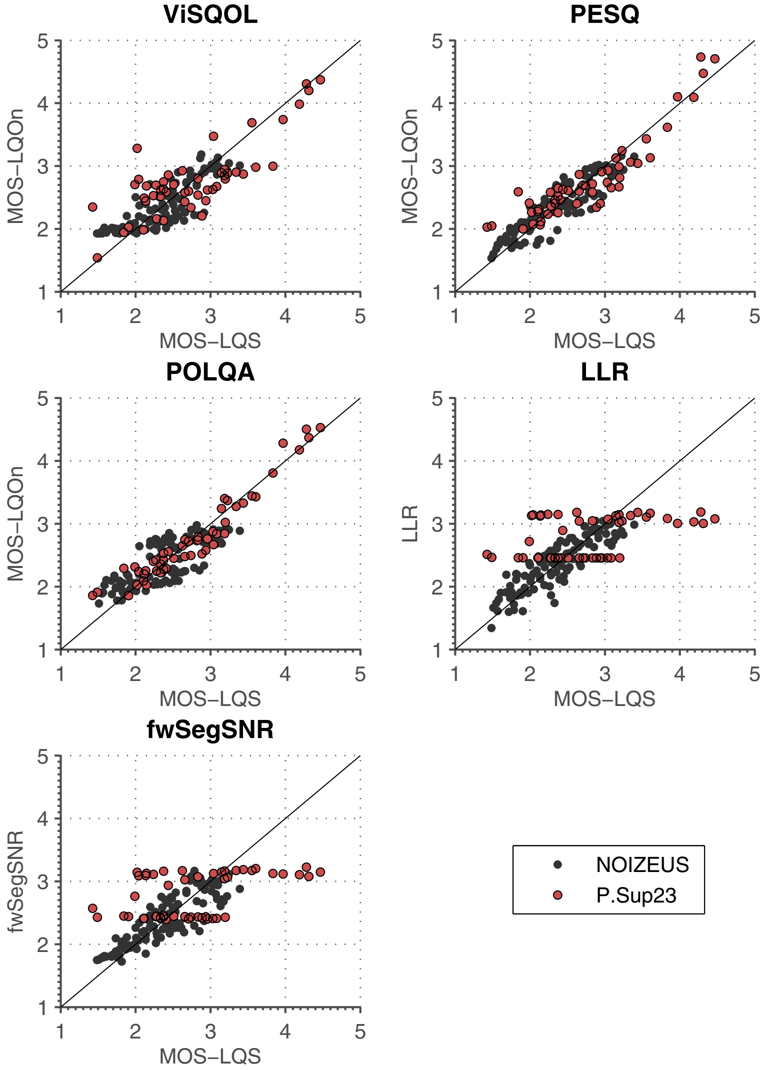

**Figure 12** Experiment 5: scatter results (NOIZEUS and P.Sup23). Objective measures against subjective MOS scores for noise and other degradations using NOIZEUS and P.Sup23 Exp 3 (Lab O; English). Plotted after linear regression fitting.

## 6 Summary and general discussion

ViSQOL shows good correlation with the NOIZEUS database subjective listener scores. The results demonstrate ViSQOL's ability to estimate speech quality in a range of background noises and also for a range of speech enhancement conditions. The P.Sup23 tests results for ViSQOL were noticeably poorer than for the other datasets, particularly in terms of the rank correlation and standard error where both PESQ and POLQA perform significantly better. Looking at the scatter plot for ViSQOL in Figure 12, the problem appears to be for lower quality samples in the MOS range of 2 to 3 where it fails to differentiate between more severe quality degradations. This may be due to the flat region in the mapping function where the raw NSIM results are tightly clustered.

For comparison, POLQA and PESQ were tested with the same test material. The results for tests with the NOIZEUS database are consistent with the performance of PESQ reported by various other authors [29,51]. Somewhat surprisingly, POLQA did not perform as well as ViSQOL or PESQ. Examining the scatter plot for POLQA in Figure 12, the NOIZEUS conditions can be seen to cluster into two groups, with a gap in the range 2 to 2.2 on the y-axis (MOS-LQOn). Further investigation showed that this gap was not a distinction based on condition, noise type, or SNR.

The Pearson correlation between all three models and the subjective quality scores were similar for the GIPS E4 database. These results had more variability within conditions, and the confidence intervals were larger than for the conditions tested in the NOIZEUS database. However, ViSQOL performed better in Spearman ranking correlation than either PESQ or POLQA for the GIPS E4 and NOIZEUS databases. The conformance test results carried out during the development of POLQA show that POLQA performs better than PESQ for all of the development and test conditions [2] tested during POLQA development. The results reported here show POLQA performed better than PESQ for the GIPS E4 tests in Experiment 4 but not the NOIZEUS tests in Experiment 5.

The correlation between subjective listener tests and objective predictions for all three models demonstrate an ability to predict subjective MOS scores when evaluated with unseen test corpora (Table 2). The PESQ model performed poorly in Experiment 1 testing warping (Figure 8). POLQA has addressed this design problem and predicts no degradation in perceived quality for up to 5% warping. ViSQOL deals with warping in a more gradual way than POLQA which is more in line with listener quality perceptions. For small, varied, imperceptible playout adjustments, ViSQOL and POLQA perform better than PESQ which shows a strong susceptibility to temporal alignment mismatches (Figure 9). For certain playout delay conditions, both ViSQOL and POLQA have shortcomings that were highlighted. ViSQOL can vary by up to 0.5 MOS for a range of adjustments and POLQA can by up to 2.5 MOS in certain conditions (Figure 10).

Overall, ViSQOL is a useful alternative to PESQ or POLQA as a full-reference speech quality model especially where VoIP systems are being evaluated. The algorithm design contains a number of properties that help deal with temporal and warping issues that can mask or distort the estimation of speech quality.

## 7 Conclusions

ViSQOL is a simple objective speech quality model based upon a mean of similarity comparisons between time-aligned, time-frequency representations of a reference and a degraded signal. Tests for a variety of conditions and VoIP-specific issues showed that it performed better than simple distance metrics and was comparable to the ITU standards, PESQ and POLQA, for wider datasets. Further work is planned with wideband speech corpora as well as for wider usage in general audio quality.

**Competing interests**

The authors declare that they have no competing interests.

**Acknowledgements**

Andrew Hines thanks Google, Inc. for support. Thanks also to Yi Hu for sharing the full listener test MOS results and enhanced test files for the NOIZEUS database.

**Author details**

[1]School of Computing, Dublin Institute of Technology, Kevin St, Dublin 8, Ireland. [2]Sigmedia, Department of Electronic and Electrical Engineering, Trinity College Dublin, College Green, Dublin 2, Ireland. [3]Google, Inc., 1600 Amphitheatre Parkway, CA 94043, Mountain View, USA.

Received: 22 October 2014 Accepted: 6 April 2015
Published online: 17 May 2015

**References**

1. ITU, Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862 (2001)
2. ITU, Perceptual objective listening quality assessment. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.863 (2011)
3. ITU, Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862.2 (2005)
4. A Hines, J Skoglund, A Kokaram, N Harte, in *Acoustic Echo Noise Control (IWAENC), IEEE Intl. Workshop on*. VISQOL: The Virtual Speech Quality Objective Listener (IEEE Aachen, Germany, 2012), pp. 1–4
5. A Hines, J Skoglund, A Kokaram, N Harte, in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA (IEEE Vancouver, Canada, 2013), pp. 3697–3701
6. H Levy, H Zlatokrilov, The effect of packet dispersion on voice applications in IP networks. IEEE/ACM Trans. Netw. **14**(2), 277–288 (2006)
7. ITU, ITU-T One-way transmission time. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.114 (2003)
8. BH Kim, H-G Kim, J Jeong, JY Kim, VoIP receiver-based adaptive playout scheduling and packet loss concealment technique. IEEE Trans. Consum. Electron. **59**(1), 250–258 (2013)
9. WebRTC, WebRTC FAQ. (http://www.webrtc.org/)
10. WebRTC, WebRTC FAQ. (http://www.webrtc.org/architecture)
11. ITU, ITU-T Methods for subjective determination of transmission quality. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.800 (1996)
12. S Möller, R Heusdens, Objective estimation of speech quality for communication systems. Proc. of the IEEE. **101**, 1955–1967 (2013)
13. T Yamada, M Kumakura, N Kitawaki, Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice. IEEE Trans. Audio Speech Lang. Process. **14**(6), 2006–2013 (2006)
14. S Möller, W-Y Chan, Côté, TH Falk, A Raake, M Waltermann, Speech quality estimation: models and trends. IEEE Signal Process. Mag. **28**(6), 18–28 (2011)
15. ITU, The E-model, a computational model for use in transmission planning (2009)
16. ITU, Wideband E-model. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.107.1 (2011)
17. ANSI ATIS, 0100005-2006: Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality (2006)
18. ITU, Single-ended method for objective speech quality assessment in narrow-band telephony applications. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.563 (2011)
19. L Sun, EC Ifeachor, Voice quality prediction models and their application in voip networks. IEEE Trans. Multimedia. **8**(4), 809–820 (2006)
20. TH Falk, W-Y Chan, Nonintrusive speech quality estimation using gaussian mixture models. IEEE Signal Process. Lett. **13**(2), 108–111 (2006)
21. V Grancharov, DY Zhao, J Lindblom, WB Kleijn, Low-complexity, nonintrusive speech quality assessment. IEEE Trans. Audio Speech Lang. Process. **14**(6), 1948–1956 (2006)
22. D Sharma, PA Naylor, ND Gaubitch, M Brookes, in *Proc. of the 19th European Signal Processing Conference (EUSIPCO)*. Short-time objective assessment of speech quality (EURASIP Barcelona, Spain, 2011), pp. 471–475
23. Z Qiao, L Sun, E Ifeachor, in *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium On*. Case study of PESQ performance in live wireless mobile voip environment (IEEE Cannes, France, 2008), pp. 1–6

24. O Slavata, J Holub, Evaluation of objective speech transmission quality measurements in packet-based networks. Comput. Stand. Interfaces. **36**, 626–630 (2014)
25. JG Beerends, C Schmidmer, J Berger, M Obermann, R Ullmann, J Pomy, M Keyhl, Perceptual Objective Listening Quality Assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement Part I – temporal alignment. J. Audio Eng. Soc. **61**(6), 366–384 (2013)
26. M Voznak, J Rozhon, P Partila, J Safarik, M Mikulec, M Mehic, *Predictive Model for Determining the Quality of a Call,* Proc. SPIE, vol. 9118, (Baltimore, Maryland, USA, 2014)
27. H Assem, D Malone, J Dunne, P O'Sullivan, in *Computing, Networking and Communications (ICNC), 2013 International Conference On*. Monitoring VoIP call quality using improved simplified e-model (IEEE San Diego, CA, USA, 2013), pp. 927–931
28. A Hines, P Pocta, H Melvin, in *Quality of Multimedia Experience (QoMEX), IEEE Workshop on*. Detailed analysis of PESQ and VISQOL behaviour in the context of playout delay adjustments introduced by voip jitter buffer algorithms (Klagenfurt am Wörthersee, Austria, 2013), pp. 18–22
29. Y Hu, PC Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(1), 229–238 (2008)
30. PC Loizou, *Speech Enhancement – Theory and Practice*. (CRC Press, Boca Raton USA, 2007)
31. A Rix, in *Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN'03)*. Comparison between subjective listening quality and P.862 PESQ score (Prague, Czech Republic, 2003)
32. A Hines, N Harte, Speech intelligibility from image processing. Speech Commun. **52**(9), 736–752 (2010)
33. A Hines, N Harte, Speech intelligibility prediction using a neurogram similarity index measure. Speech Commun. **54**(2), 306–320 (2012)
34. MSA Zilany, IC Bruce, PC Nelson, LH Carney, A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. J. Acoust. Soc. Am. **126**(5), 2390–2412 (2009)
35. Z Wang, AC Bovik, HR Sheikh, EP Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
36. S Kandadai, J Hardin, CD Creusere, in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. Audio quality assessment using the mean structural similarity measure (IEEE Las Vegas, NV, USA, 2008), pp. 221–224
37. D Breakey, C Meskell, Comparison of metrics for the evaluation of similarity in acoustic pressure signals. J. Sound Vibration. **332**(15), 3605–3609 (2013)
38. W Voiers, in *Acoustics, Speech, and Signal Processing, 1980. ICASSP 1980. IEEE International Conference on*. Interdependencies among measures of speech intelligibility and speech "quality", vol. 5 (IEEE Denver, CO, USA, 1980), pp. 703–705
39. A Hines, N Harte, in *Speech Perception and Auditory Disorders*, ed. by T. Dau, et al. Comparing hearing aid algorithm performance using simulated performance intensity functions (Danavox Jubilee Foundation Ballerup, Denmark, 2011), pp. 347–354
40. ViSQOL Software. http://sigmedia.tv/tools
41. ANSI, *ANSI S3.5-1997 (R2007). Methods for Calculation of the Speech Intelligibility Index*. (American National Standards Institute, New York, 1997)
42. P Pocta, H Melvin, A Hines, An analysis of the impact of playout delay adjustments introduced by VoIP jitter buffers on speech quality. Acta Acoust. united Acustica. **101**(3), 616–631 (2015)
43. IEEE, IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoustics. **17**(3), 225–246 (1969)
44. ITU, ITU-T coded-speech database. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.Sup23 (1998)
45. C Hoene, H Karl, A Wolisz, A perceptual quality model intended for adaptive VoIP applications. Int. J. Commun. Syst. **19**(3), 299–316 (2005)
46. WebRTC, WebRTC FAQ. (http://www.webrtc.org/faq#TOC-What-is-the-iSAC-audio-codec-)
47. ITU, Test signals for use in telephonometry, Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.501 (2012)
48. M-K Lee, H-G Kang, Speech quality estimation of voice over internet protocol codec using a packet loss impairment model. J. Acoust. Soc. Am. **134**(5), 438–444 (2013)
49. Y Hu, PC Loizou, Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. **49**(7-8), 588–601 (2007)
50. ITU, Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.835 (2003)
51. AA Kressner, DV Anderson, CJ Rozell, in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. Robustness of the hearing aid speech quality index (HASQI) (IEEE New Paltz, NY, USA, 2011), pp. 209–212