VISQOL: THE VIRTUAL SPEECH QUALITY OBJECTIVE LISTENER

Andrew Hines,[‡*] *Jan Skoglund,*[†] *Anil Kokaram* [†] *and Naomi Harte*[‡]

[‡] Sigmedia, Trinity College Dublin, Ireland [†] Google, Inc., Mountain View, CA, USA

ABSTRACT

A model of human speech quality perception has been developed to provide an objective measure for predicting subjective quality assessments. The Virtual Speech Quality Objective Listener (ViSQOL) model is a signal based full reference metric that uses a spectro-temporal measure of similarity between a reference and a test speech signal. This paper describes the algorithm and compares the results with PESQ for common problems in VoIP: clock drift, associated time warping and jitter. The results indicate that ViSQOL is less prone to underestimation of speech quality in both scenarios than the ITU standard.

1. INTRODUCTION

Perceptual measures of quality of experience rather than quality of service are becoming more important as transmission channels for human speech communication have evolved from a dominance of POTS to a greater reliance on VoIP. Accurate reproduction of the input signal is less important, as long as the user perceives the output signal as a high quality representation of the original input.

PESQ (Perceptual Evaluation of Speech Quality) [1] and its recent successor POLQA (Perceptual Objective Listening Quality Assessment) [2] are full reference measures described in ITU standards that 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. The newer POLQA model yields quality estimates for both narrowband and super-wideband speech and addresses other limitations in PESQ. It is not yet in widespread use or freely available for testing and has not been evaluated in this study.

NSIM (Neurogram Similarity Index Measure) was originally developed as a full-reference measure for predicting speech intelligibility [3]. This paper adapts the NSIM methodology to the domain of speech quality prediction. We concentrate specifically on areas of speech quality assessment where PESQ has known weaknesses. Clock drift is a commonly encountered problem in VoIP systems which can cause a drop in speech quality estimates from PESQ, but in reality does not impact on the user perception of speech quality. Small resulting changes, such as some temporal or frequency warping, may be imperceptible to the human ear and should not necessarily be judged as a quality degradation. Jitter may not always be fully corrected in cases where the jitter buffer is not sufficiently long, even with no packet loss. This can cause the received signal to be speeded up or slowed down to maintain overall delay, an effect that will not impact overall perceived quality in a call when low enough.

This paper presents an analysis of the use of NSIM as the basis of the development of a Virtual Speech Quality Objective Listener (ViSQOL) model. Realistic examples of time warping and jitter are assessed for speech quality using PESQ and the results compared to the newly developed ViSQOL. Section 2 gives further background on the measures of PESQ and NSIM. Section 3 describes the VisQOL model architecture and sections 4 and 5 introduce the experiments involving clock drift and jitter typical of modern VoIP communications. The discussion in section 6 highlights the ViSQOL model's ability to predict and estimate time warping and discusses its further potential.

[*]Thanks to Google, Inc. for funding. Email: andrew.hines@tcd.ie

2. QUALITY MEASURES

2.1. PESQ

PESQ is a full reference metric that compares two signals before and after passing through a communications channel to predict speech quality. The signals are time aligned, followed by a quality calculation based on a psychophysical representation. Quality is scored in a range of -0.5 to 4.5, although the results for speech are usually in the range of 1 to 4.5. A transfer function mapping from PESQ to MOS has been developed using a large speech corpus [4]. The original PESQ metric was developed for use on narrowband signals (300-3,400 Hz). It deals with a range of transmission channel problems including speech input levels, multiple bit rate mode codecs, varying delays, packet loss and environmental noise at the transmission side. It is acknowledged in the ITU standard that PESQ provides inaccurate predictions for quality involving a number of other issues: listening levels, time warping, loudness loss, effects of delay in conversational tests, talker echo and side tones. PESQ has evolved over the last decade with a number of extensions.

2.2. NSIM

The Neurogram Similarity Index Measure (NSIM) [3] was developed by the authors to evaluate the auditory nerve discharge outputs of models simulating the working of the ear. A neurogram is analogous to a spectrogram with colour intensity related to neural firing activity. NSIM rates the similarity of neurograms and can be used as a full reference metric to predict speech intelligibility. While speech intelligibility and speech quality are linked, work by Voiers [5] showed that an amplitude distorted signal that had been peak-clipped did not seriously affect intelligibility but seriously affected the aesthetic quality. 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 [6]. 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 modelling 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.

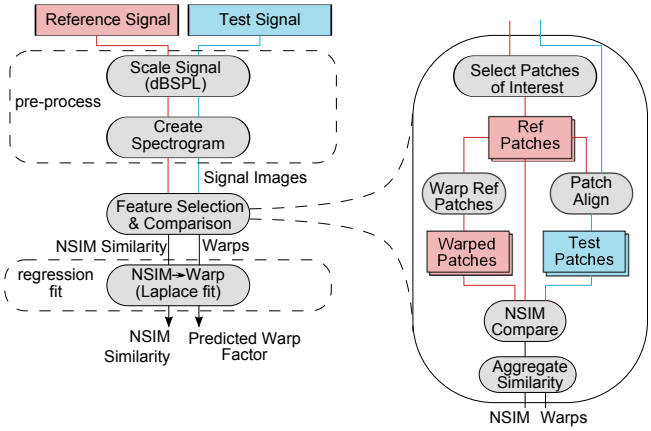

Fig. 1. Flow diagram for ViSQOL.

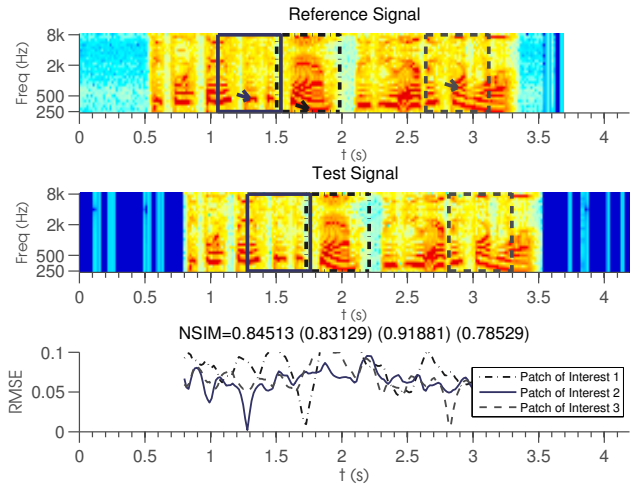

Fig. 2. Jitter Signal Example. The spectrogram of the original signal is shown above the degraded. The patch windows are shown on both signals with a small pointer in the centre of the reference windows showing the frequency band used to select the patch of interest. Each patch is 30 frames. The RMSE correlation shown in the bottom pane also illustrates how the patches in the degraded signal were aligned to the reference patches. The mean NSIM for the three patches is shown with the NSIM per patch in parenthesis.

3. VISQOL MODEL ARCHITECTURE

ViSQOL is a model of human sensitivity to degradations in speech quality. It compares a reference signal with a degraded test signal. The output is a prediction of speech quality perceived by an average individual. The model has three major processing stages shown in Fig. 1: pre-processing, feature selection and comparison, and a regression fitted transfer function. The pre-processing stage scales the test signal to match the reference signal's sound pressure level. Short-term Fourier Transform (STFT) spectrogram representations of the reference and test signals are created with 30 frequency bands logarithmically spaced between 250 and 8,000 Hz. A 512 sample, 50% overlap 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. The spectrograms are used as inputs to the second stage of the model, shown in detail on the right-hand side of Fig. 1.

The aim of the *feature selection and comparison* stage is to identify corresponding patches in the reference and degraded spectrograms. Three patches are selected from the reference signal for comparison, each 30 frames long by 30 frequency bands (23 bands, i.e. 250-3,400 Hz, are used for narrowband quality assessment). The patches are automatically chosen by finding the maximum intensity frame in each of three frequency bands (bands 2, 6 and 10 corresponding roughly to 250, 450 and 750 Hz. These points are marked with a small arrow in the middle of the reference patch boxes in the Fig. 2 example.) This mechanism ensures that the patches of interest contain speech content rather than silences and are likely contain structured vowel phonemes with strongly comparative features. While patches can potentially overlap there is generally a good spread between them.

Each reference patch is aligned with the corresponding area from the test spectrogram. A relative mean squared error (RMSE) difference is carried out between the reference patch and a test spectrogram patch frame by frame, thus identifying the maximum correlation point for each patch. The bottom pane in Fig. 2 shows the RMSE for each patch with the matched patches marked on the test spectrogram at their RMSE minima. RMSE is only used for patch alignment as it is an unbounded metric. NSIM is used to predict the similarity and quality of the aligned patches.

NSIM is more sensitive to time warping than a human listener. The model counteracts this by warping the spectrogram patches temporally. It creates alternative reference patches from 1% to 5% longer and shorter than the original reference. The patches are created using a cubic two-dimensional interpolation. The comparison stage is completed by comparing the test patches to both the reference patches and 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. The mean NSIM score for the three test patches is returned as the signal similarity estimate. NSIM outputs a bounded score between 0 and 1 for the range from no similarity to identical. A final stage uses a Laplacian function fitted to training data to predict the amount of time warping in the test signal.

4. EXPERIMENT 1: CLOCK DRIFT SIMULATION

This experiment simulates time warp distortion of signals due to low frequency clock drift between the signal transmitter and receiver. Clock drift can cause delay problems if not detected and seriously impact VoIP conversation quality, but a small drift of (e.g. 1 to 4 or 5%) is not noticeable to a listener when comparing over a short speech sample. Clock drift can be mitigated using clock synchronisation algorithms at a network level by analysing packet time-stamps but the clock drift can be masked by other factors such as jitter when packets arrive out of synchronisation.

Ten sentences from the IEEE Harvard Speech Corpus were used as reference speech signals [7]. The 8 kHz sampled reference signals were originally resampled to create time warped versions. The reference and resampled test signal were evaluated with both PESQ and the ViSQOL model. The test was repeated for reference signals with a range of resampled test signals, with resampling factors ranging from 0.85 to 1.15.

The results are presented in Fig. 3. The mean speech quality predictions are plotted against the resampling factor, with error bars showing the standard deviation. The top plot shows the PESQ results, the middle shows the ViSQOL model results (NSIM is the scale unit) and the bottom plot shows a stack bar breakdown of the warped patches used by chosen by ViSQOL for the similarity measure. Looking at the comparison between the PESQ and ViSQOL models, it is evident that the full ranges of both metrics are covered

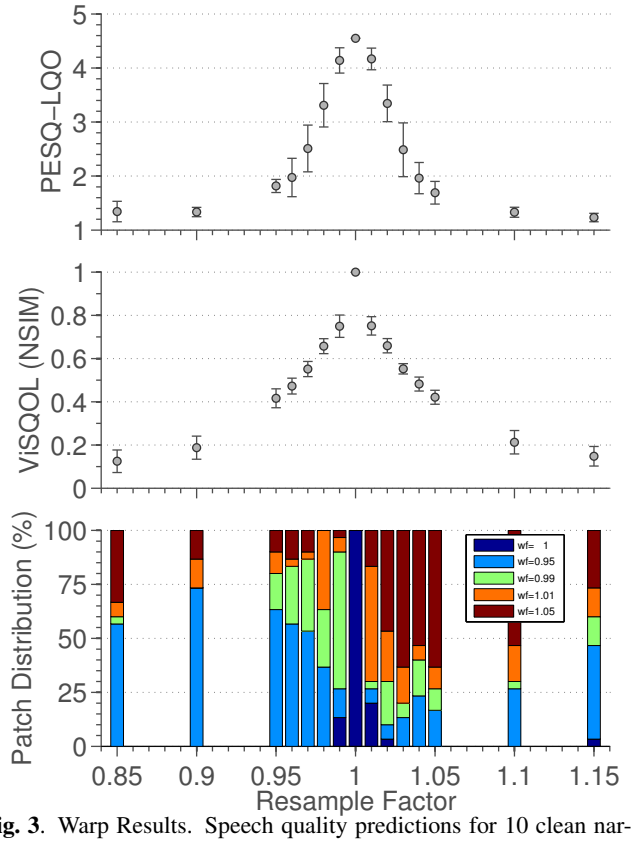

Fig. 3. Warp Results. Speech quality predictions for 10 clean narrowband sentences. Top two plots: PESQ and ViSQOL speech quality predictions showing mean values at each resampling factor compared to the reference signals. Error bars are standard deviation. Bottom: Distribution plot of warped patch sizes used for each signal resampling factor. WF in legend refers to the patch warp factor.

by the test. Both follow a similar trend with plateaus at the extremities and symmetry around the non-resampled perfect quality comparison maximum. It should be noted that prior experience measuring speech similarity with NSIM [3] found a practical peak similarity even for small differences at approximately 0.8. Hence the fall off from the reference comparison at 1.0 is not as steep as this graph might suggest. Listening to the resampled tests, the differences are not audible at 2% resampling or less. Although a change in pitch is noticeable, the change is not a dramatic degradation in quality until 5% to 10%. The PESQ predictions show a dramatic drop in predicted quality between 3% and 4% resampling whereas the NSIM drop occurs later between 5% and 10%, which matches the listener experience. The standard deviation for PESQ is significantly larger than for ViSQOL which is more consistent for the same time warp.

The stacked bar plot under the ViSQOL results illustrates the distribution of warped reference patch usage by ViSQOL in calculating the NSIM similarity. The y-axis shows the number of patches for each patch warp factor that were used with signals of a given resampling. The model uses the maximum similarity from the test patch compared with the reference patch and its warped reference patches. As the resampling increases, so the warp factor of the selected patches increases. As expected, the patch distribution shows that the non-resampled reference only uses unwarped patches and the reliance on larger warps grows as the resampling increases. However, less intuitively, the warp factors do not necessarily match exactly with the resampling factors. The NSIM scores combined with

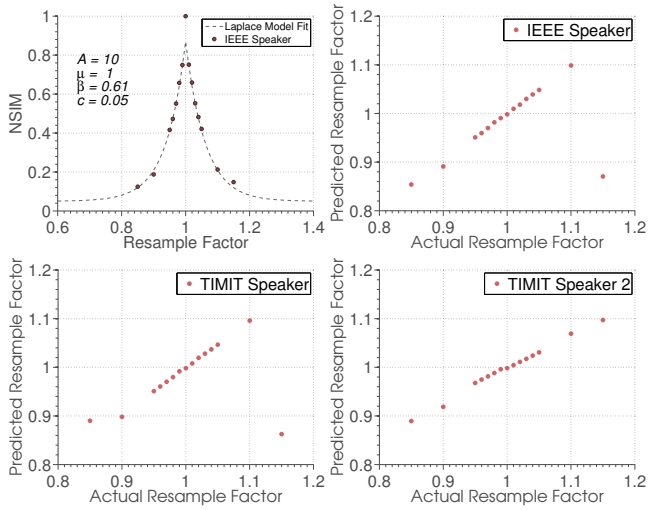

Fig. 4. Top Left: Model Fit of Laplace function to *IEEE Speaker* data. Top Right and Bottom: Mean predicted warp for 10 samples for 3 test speakers.

knowledge of the warped patches used is discussed below where a potential application of ViSQOL in the detection of clockdrift above the network layer is presented.

4.1. Predicting Time Warping

The ViSQOL output can be used to predict time warping in speech samples by fitting a regression model to the NSIM data. A Laplacian function,

$$y = \frac{e^{\frac{-A|x-\mu|}{\beta}}}{2\beta} + c \tag{1}$$

was fitted to the mean NSIM scores for each resample factor. The fitted function is shown in Fig. 4. By inverting (1), a function for predicting the warp factor for a given NSIM can be obtained as

$$x = \frac{b}{A}\ln(2b(y-c)) + \mu, \qquad 0.06 \geq y \geq 0.89. \tag{2}$$

The symmetrical nature of the function means that it will not predict whether the test signal's resample factor is greater or less than the reference signal. To determine which side of the Laplacian slope should be predicted, the warp factors used in the patches are examined. The ratio of patches smaller than the original size versus those larger than the original size and the resample factor prediction is adjusted to match.

Fig. 4 shows the results for the *IEEE Speaker* from experiment 1 which was used to obtain the model fit as well as two other test sets: *TIMIT*[1] *Speaker* and *TIMIT Speaker 2*, a female and male speaker. Each test featured a single speaker and 10 reference sentences with 14 warp factors per sentence. The scatter diagrams show the actual resample factor plotted on the x-axis against the predicted resample factor on the y-axis. The points are mean predicted values for the 10 sentences. It is clear from the results that the model is very accurate at predicting warps of 10% around the reference rate for clean data. The magnitude of warps at 15% are still predicted well but the model failed in both the *IEEE Speaker* and *TIMIT Speaker* cases to detect whether it is a higher or lower sampling rate detected, resulting in a warp factor of 1.15 being predicted as 0.85.

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

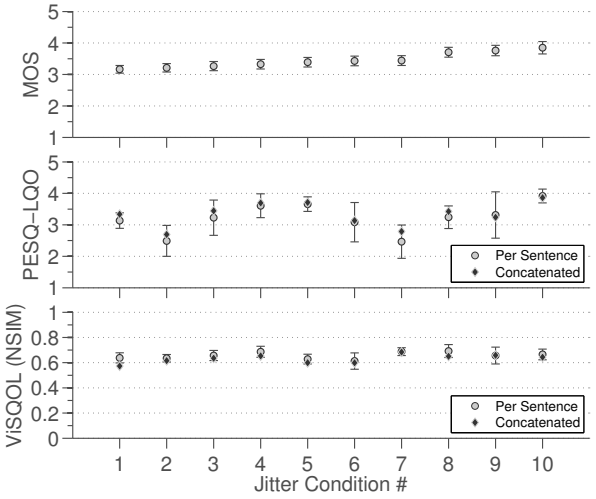

Fig. 5. Speech quality predictions for 10 jitter conditions. Top: MOS scores and 95% CI for actual listeners listening to 8 sentences in each condition. Middle and Bottom PESQ (MOS-LQO) and ViSQOL speech quality predictions. Circles and 95% CI show results for testing each sentence per condition. Diamonds show results for testing against concatenated sentences.

5. EXPERIMENT 2: JITTER

A second experiment used 8 IEEE sentences: 2 sentences spoken by 2 male and 2 female speakers. These were concatenated to form a reference signal. Ten jitter test conditions containing varying amounts of non-uniform warping were created from the reference signal. MOS scores based on 25 listeners with 4 votes per condition are shown in Fig. 5 along with corresponding results for PESQ and the ViSQOL model. For PESQ and ViSQOL, quality was estimated both for the complete concatenated 8 sentences (32 seconds) and per sentence (4 seconds). The diamonds show the predicted scores based on the full signal comparisons and the circles show the estimates per sentence with 95% confidence intervals (CIs). The mean scores *per sentence* and the *concatenated* scores correspond closely for both metrics, however the 95% CI for the MOS-LQO per sentence show the wide variability between sentences in the same condition. This is not a feature in the ViSQOL results. The results for the 10 conditions are presented in order of ascending MOS scores and range from 3.2 to 3.9. The MOS-LQO results do not predict the same ranking trend as the MOS scores and have a wider range between 2.4 and 3.9. As for PESQ, ViSQOL does not rank the conditions in the same order as the MOS test, but the range, with NSIM scores between 0.61 and 0.69 and much smaller error bars shows a better ability to handle jitter and predict the impact on quality in a consistent manner. This result is a promising indicator of ViSQOL's ability to provide consistent quality measures in varying jitter conditions, even over short periods. The next step in the process of establishing ViSQOL as a quality measure is to make the connection between the NSIM score and the MOS score. This will require a comprehensive evaluation of the metric. The initial attractive attributes of this metric are demonstrated in these two experiment: an ability to detect clock drift; and consistent sensitivity to jitter.

6. DISCUSSION AND FUTURE WORK

The results demonstrate the ViSQOL model's ability to detect and quantify clock drift and jitter. The tests focused on detecting constant and varying time warping. Based on short speech samples, temporally varying warps are handled more consistently by ViSQOL than PESQ. This is a useful property as whilst there are a range of QoS metrics available to predict delay and clock drift, their ability to predict the end user perceptual quality of experience is limited [8]. The experimental results highlighted the large deviation in predicted quality exhibited by PESQ for small sampling factor changes, and for short samples of variable warping.

The model is still in the early stages of development and while the results are promising there are a range of issues requiring further analysis. The key decisions in the evolution of the model's parameters included evaluating and testing: the number of patches; the frequency bands used to determine the patch locations in the reference signal; and the number of warp factors to be evaluated. The optimal values were chosen and used in the experiments presented.

This work focused on narrowband signals but the model is open to adaptation by adjusting the parameters of the spectrogram images to suit the wideband signals commonly used in VoIP. ViSQOL was developed as a full objective speech quality prediction tool and further work is underway to develop a transfer function to map the NSIM output from the model to a predicted MOS score. The current model could also be used in combination with PESQ to flag poor quality estimates caused by time warping.

This paper has introduced ViSQOL as a model for predicting speech quality. Specifically, the ability to detect and predict the level of clock drift or jitter and whether it will impact a listeners quality of experience was investigated. It was shown that ViSQOL can detect clock drift and jitter and also predict the magnitude of clock drift distortion.

7. 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] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Commun.*, vol. 54, no. 2, pp. 306 – 320, 2012.

[4] ITU, "Mapping function for transforming P.862 raw result scores to MOS-LQO," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862.1, 2003.

[5] W. Voiers, "Interdependencies among measures of speech intelligility and speech "quality"," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.*, 1980, vol. 5, pp. 703–705.

[6] A. Hines and N. Harte, "Comparing hearing aid algorithm performance using simulated performance intensity functions," in *Speech Perception and Auditory Disorders (494 pp)*, T. Dau et al., Ed., pp. 347–354. Danavox Jubilee Foundation, 2011.

[7] IEEE, "IEEE recommended practice for speech quality measurements," *Audio and Electroacoustics, IEEE Transactions on*, vol. 17, no. 3, pp. 225–246, Sep 1969.

[8] W. Jiang and H. Schulzrinne, "QoS measurement of internet real-time multimedia services," Technical report, Columbia University, 1999.