ROBUSTNESS OF SPEECH QUALITY METRICS TO BACKGROUND NOISE AND NETWORK DEGRADATIONS: COMPARING VISQOL, PESQ AND POLQA

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

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

ABSTRACT

The Virtual Speech Quality Objective Listener (ViSQOL) is a new objective speech quality model. It is a signal based full reference metric that uses a spectro-temporal measure of similarity between a reference and a test speech signal. ViSQOL aims to predict the overall quality of experience for the end listener whether the cause of speech quality degradation is due to ambient noise, or transmission channel degradations. This paper describes the algorithm and tests the model using two speech corpora: NOIZEUS and E4. The NOIZEUS corpus contains speech under a variety of background noise types, speech enhancement methods, and SNR levels. The E4 corpus contains voice over IP degradations including packet loss, jitter and clock drift. The results are compared with the ITU-T objective models for speech quality: PESQ and POLQA. The behaviour of the metrics are also evaluated under simulated time warp conditions. The results show that for both datasets ViSQOL performed comparably with PESQ. POLQA was shown to have lower correlation with subjective scores than the other metrics for the NOIZEUS database.

Index Terms— Objective Speech Quality, POLQA, P.853, ViSQOL, NSIM

1. INTRODUCTION

Accurately predicting a listener's perceptual rating of speech quality using subjective testing is an active topic of research and has resulted in a number of industry standards [1, 2]. Perceptual measures of quality of experience are continuously evolving as the variety of communication channels for human speech communication has grown. From an original dominance of narrowband telephony, the range of channels has expanded. Multimedia conferencing such as Google Hangouts and Skype have increased the popularity and usage of voice over internet protocol (VoIP) for video conferencing. The content is delivered using a standard computer or mobile device rather than dedicated video conferencing hardware. End-to-end, the speech delivery channel has become more complex and the number of variables impacting quality of experience has expanded. These changes have altered the scope for objective measures but the goal remains the same: predicting listeners' subjective opinion of quality. It is important to keep this in mind when developing metrics for speech quality prediction as speech codecs now seek to target perceptual fidelity to the input signal rather than signal reproduction fidelity itself. The user must perceive 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, so there has been limited publication of its performance outside of its own development and conformance tests.

The Virtual Speech Quality Objective Listener, ViSQOL, is a full reference objective speech quality model of human sensitivity to degradations. Prior work [3] demonstrated the ViSQOL model's ability to detect and quantify clock drift and jitter for VoIP transmission. 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.

The aim of this work was to establish ViSQOL's ability to predict speech quality using a wider ranging set of test data with speech degraded under a variety of conditions. Two speech corpora were used, which focus on different kinds of degradations. The NOIZEUS corpus contains speech under a variety of background noises while the GIPS Exp. 4 (E4) database contains degradations encountered in VoIP. This work builds on prior tests where PESQ was tested against the NOIZEUS database [4, 5] and POLQA has been tested against the E4 database in a wideband scenario [2]. Here, the robustness of the ViSQOL model is benchmarked against PESQ and POLQA.

A second experiment compared the ability of the three models to handle time warping in speech signals. 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.

Section 2 gives further background on the PESQ, POLQA and ViSQOL models with a detailed description of the ViSQOL model architecture in section 2.2. The experimental setup is presented in section 3 and the speech corpora used in the experiment are outlined in section 2.3. The discussion in section 6 examines the comparative results and comments on performance of ViSQOL and POLQA as well as discussing the impact of improvements to the ViSQOL model and its further potential.

These experiments represent the first benchmarking of POLQA against datasets not used during the ITU development and conformance tests for the model.

Poor handling of time warping was an acknowledged shortcoming of PESQ that POLQA sought to address. The improvements are quantified in this work and the results are also compared to the warping handling of ViSQOL which has been designed with VoIP specifically in mind.

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

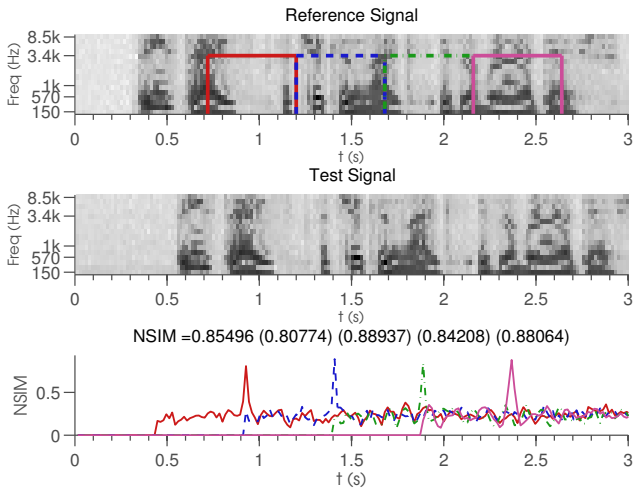

Fig. 1. Speech Signals with sample patches. The bottom plot shows the NSIM similarity score for each patch from the reference compared frame by frame across the test signal. The NSIM score is the mean of the individual patch scores given in parenthesis.

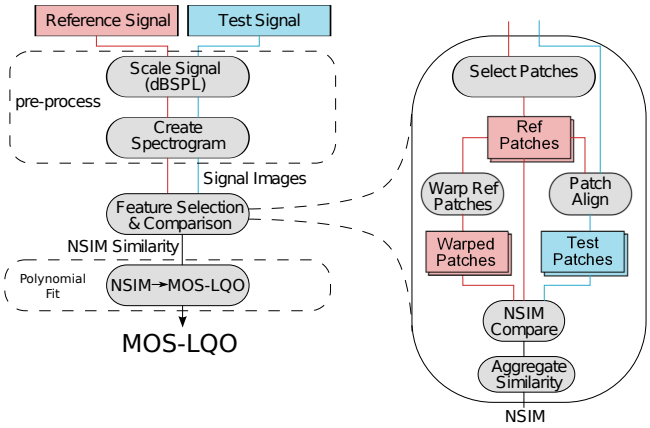

Fig. 2. Flow diagram for ViSQOL.

2. BACKGROUND: MEASURING SPEECH QUALITY

2.1. Full Reference Metrics

Subjective tests use an absolute category rating (ACR) from 1 to 5 to score speech quality. They are time consuming and expensive and accurately predicting mean opinion scores (MOS) objectively has been an active research topic for a number of decades. Full reference tests compare a degraded signal to a clean original to predict the speech quality. The two signals are time aligned, followed by a quality calculation based on a psychophysical representation. The ITU-T recommended standard 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 last decade. It was originally designed and tested on narrowband signals (300-3,400 Hz). 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]. POLQA is the ITU-T P.863 standard that supersedes PESQ and addresses a range of the limitations of PESQ as well as improving the overall correlation with subjective MOS scores. It allows for predicting overall listening speech quality in two modes: narrowband (300 to 3,400 Hz) and superwideband (50 to 14,000 Hz). The tests described below were carried out using POLQA 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.

2.2. ViSQOL

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. 2: pre-processing, alignment and comparison. 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 using critical bands between 150 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 test spectrograms are floored to the minimum value in the reference spectrogram to ensure that negative values in the test spectrogram caused by packet loss do not distort the similarity scores. The spectrograms are used as inputs to the second stage of the model, shown in detail on the right-hand side of Fig. 2.

The reference signal is segmented into patches for comparison as illustrated in Fig. 1. Each patch is 30 frames long by 16 critical frequency bands [6] (i.e. 150-3,400 Hz). ViSQOL can be configured to use more bands (21 bands up to 8kHz) if the signals to be tested are wideband. Each reference patch is aligned with the corresponding area from the test spectrogram. The Neurogram Similarity Index Measure (NSIM) [7] is used to measure the similarity between the reference patch and a test spectrogram patch frame by frame, thus identifying the maximum correlation point for each patch. This is shown in the bottom pane of Fig. 1 where each line graphs the NSIM based correlation function for each patch in the reference signal compared with the example signal. The NSIM at the maxima is averaged over the patches to yield the metric for the example signal.

NSIM was originally developed by the authors to evaluate the auditory nerve discharge outputs of models simulating the working of the ear. It was inspired by the image quality metric SSIM [8] but adapted and developed for use in the auditory domain. NSIM is more sensitive to time warping than a human listener. The ViSQOL 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 test patches is returned as the signal similarity estimate. Finally, a 3rd order polynomial fitting function is used to translate the NSIM similarity score into a MOS-LQOn score and mapped in the range -0.5 to 4.5.

ViSQOL was introduced in prior work [3] but has been further developed and adjusted to ensure accurate estimation of some conditions, specifically packet loss. For this work there were three important changes made to the model. The original version only

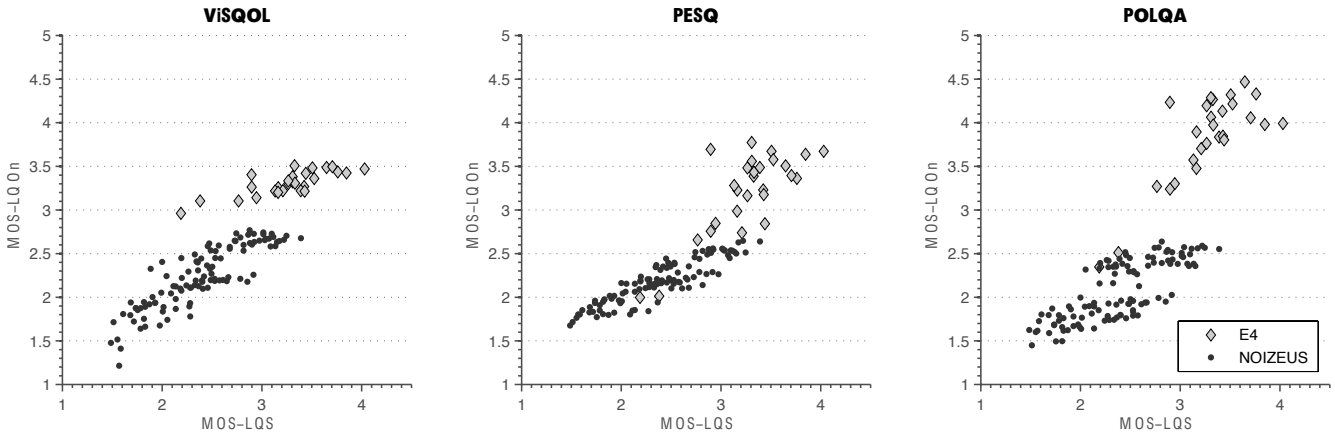

Fig. 3. Scatter Plot Results for ViSQOL, POLQA and PESQ.

selected 3 patches from the reference speech spectrogram and this worked well where the degradation was uniform across the test sample. For intermittent degradations, e.g. packet loss, there was a high chance that the degradation may be missed altogether. This was easily recognised as an issue as for packet loss conditions, the predicted MOS-LQOn were noticeable as outliers in the scatter plots. The model was adjusted accordingly to evaluate patches covering the full voiced area of the reference speech spectrogram. A second important change was switching from RMSE to NSIM for the patch alignment stage. When testing in high levels of background noise, RMSE was failing to correctly align patches in the degraded spectrogram with the reference signal. Switching to NSIM yielded a significant improvement in accuracy. The final change was from logarithmically based frequency bands to critical bands which provided higher correlation scores.

2.3. NOIZEUS and GIPS Exp. 4 (E4) Speech Databases

NOIZEUS [9] is a narrowband 8k sampled noisy speech corpus that was originally developed for evaluation of speech enhancement algorithms. Mean opinion scores (MOS) for a subset of the corpus were obtained using the ITU-T P.835 standard methodology for subjective evaluation. Four noise types from the full corpus were tested: babble, car street and train. Each noise type was tested with 13 speech enhancement algorithms plus the 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 2 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.

A second speech quality corpus, referred to in this paper as the E4 corpus, contains tests of the wideband codec iSAC [10] with superwideband references. GIPS NetEQ was used for jitter buffer handling. The test was designed as a Mean Opinion Score (MOS) listening assessment, performed in Native British English. A sliding scale ACR test was used. Within the experiments the iSAC wideband codec was assessed, with respect to MNRU and reference speech codec/conditions. The processed sentence-pairs were each scored by 25 listeners. The sentences are from ITU-T P.501 which contains 2 male and 2 female (British) English speakers sampled at 48 kHz and were downsampled to 16 kHz for the listening test.

For these tests all signals were downsampled to 8k narrowband signals. Twenty seven conditions from the corpus were tested with 4 speakers per condition (2 male and 2 female). Twenty-five listeners scored each test sample, resulting in 100 votes per condition. The breakdown of conditions was as follows: 10 jitter, 13 packet loss and 4 clock drift.

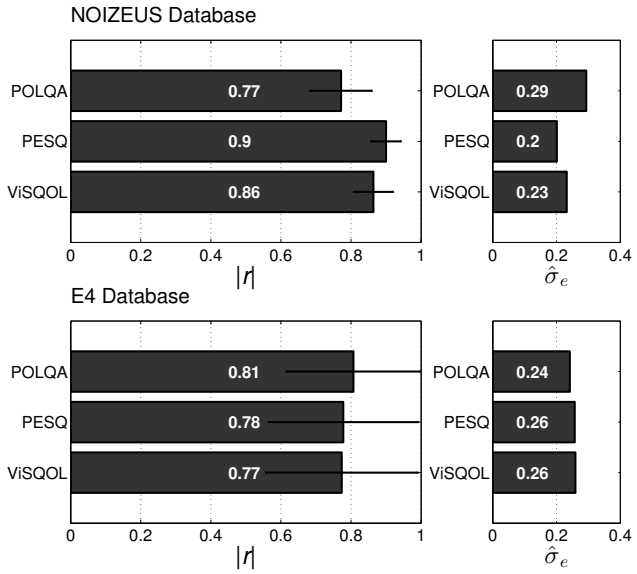

Fig. 4. Results for NOIZEUS and E4 corpora. Left: Absolute value of the correlation between subjective and objective scores with 95% confidence interval error bars. Right: Estimate of the standard deviation of the error based on the correlation coefficient.

3. EXPERIMENT 1: NOIZEUS AND E4 TESTS

Hu and Loizou [4] 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 [4]. They measured Pearson's correlation coefficient across the 112 conditions (as described in section 2.3) for each measure as well as the standard deviation

of the error. For predicting overall quality, they found PESQ generated the highest correlation of the metrics tested. Absolute values of Pearson's correlation coefficient, $|r|$, can be calculated using,

$$r = \frac{\sum_i (o_i - \bar{o})(s_i - \bar{s})}{\sqrt{\sum_i (o_i - \bar{o})^2} \sqrt{\sum_i (s_i - \bar{s})^2}} \quad (1)$$

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 $\bar{o}$ and $\bar{s}$ respectively. The standard deviation of the error, $\hat{\sigma}_e$, was also measured as a secondary test,

$$\hat{\sigma}_e = \hat{\sigma}_s \sqrt{1 - r^2} \quad (2)$$

where $\hat{\sigma}_s$ is the standard deviation of the subjective quality scores, s and r is the correlation coefficient. Hu and Loizou [4] split their data for training and testing. Subsequent evaluations by Kressner et al. [5] repeated the experiments using the full dataset of 1792 speech files, which is the approach adopted in this study.

The three objective speech quality metrics, ViSQOL, PESQ and POLQA were tested using the NOIZEUS and E4 corpora. Results were averaged by condition and compared to the average MOS scores per condition. Fig. 3 shows the results for each objective quality measure. The scatter shows 112 NOIZEUS conditions (circles) and 30 E4 conditions (diamonds). The correlation coefficients and standard deviation of the error for each corpus are presented in Fig. 4. The error bars for the correlation are 95% confidence intervals calculated using Fisher's z transformation.

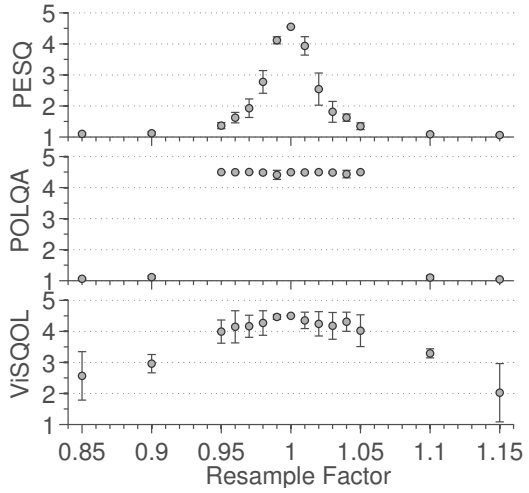

Fig. 5. Experiement 2: Warp Results. MOS-LQOn predictions for each model comparing 10 sentences for each resample factor.

4. EXPERIMENT 2: CLOCK DRIFT SIMULATION

This experiment was originally presented as a comparison between PESQ and ViSQOL [3]. Here, the experiment was extended to include POLQA and the ViSQOL results were mapped to a MOS-LQOn scale to allow the three metrics to be compared. 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. Ten sentences from the IEEE Harvard Speech Corpus were used as reference speech signals [11]. The reference and resampled test signal were evaluated with using PESQ, POLQA and ViSQOL for each sentence at each resampling factor.

5. RESULTS AND DISCUSSION

The correlation displayed by all three models demonstrated an ability to predict subjective MOS scores when evaluated with unseen test corpora. The results for tests with the NOIZEUS database are consistent with the performance of PESQ reported by various other authors [4, 5]. The results also demonstrated ViSQOL's ability to estimate speech quality in a range of background noises and also for a range of speech enhancement conditions. Somewhat surprisingly, POLQA did not perform as well as ViSQOL or PESQ. Examining the scatter plot for POLQA in Fig. 3, the NOIZEUS conditions can be seen to cluster into two groups, with a gap in the range 2–2.2 on the y-axis (MOS-LQOn). Further investigation showed that this gap was not a distinction based on condition, noise type, or SNR, hinting that it may be something to do with the mapping function used by POLQA to map the raw similarity score to MOS-LQOn.

The E4 database results had similar correlation with subjective scores for all three models. These results had more variability within conditions and the confidence intervals were larger than for the conditions tested in the NOIZEUS database. Fig. 4 presents the standard deviation of the errors which exhibit the same ranking trends as the correlation results. 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]. The results reported here show a better performance than PESQ for the E4 tests but not the NOIZEUS tests. The scatter of E4 conditions highlights that ViSQOL tended to under-predict the MOS scores for the E4 conditions and that they were more tightly clustered than the PESQ and POLQA predictions.

The second experiment tested the robustness of the three models to time warping. Fig. 5 presents the results the resample factors from .85 to 1.15 along the x-axis against MOS-LQOn quality predictions for the 3 metrics. The PESQ model performs poorly, predicting a drop in quality for factors of 1–4% where the average listener perceives no difference. The newer POLQA standard addressed this problem and predicts no degradation in perceived quality for up to 5% warping. However it drops off steeply and predicts a MOS-LQOn of 1 for warping of 10%. Warping of this size causes 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 a listeners' opinion.

6. CONCLUSIONS AND FUTURE WORK

ViSQOL is a simple objective speech quality model that has not been trained with any datasets in its quality prediction. It relies solely upon a similarity comparison between time-frequency representations of a clean and a distorted signal. The test results presented here show that it has comparable results to PESQ and POLQA for two speech corpora, NOIZEUS and E4, which contain very different quality degradations. Work is currently underway to test with a wide range of other speech databases and testing with wideband speech corpora would allow benchmarking with POLQA for VoIP scenarios where wideband or superwideband speech is used.

7. ACKNOWLEDGEMENTS

Thanks to Yi Hu for sharing the full listener test MOS results and enhanced test files for the NOIZEUS database.

8. 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, J. Skoglund, A. Kokaram, and N. Harte, "VISQOL: The Virtual Speech Quality Objective Listener," in *IWAENC*, 2012.

[4] Yi Hu and Philopos C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.

[5] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Robustness of the hearing aid speech quality index (HASQI)," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, 2011, pp. 209–212.

[6] ANSI, *ANSI S3.5-1997 (R2007). Methods for calculation of the speech intelligibility index.*, American National Standards Institute, 1997.

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

[8] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[9] Yi Hu and Philipos C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[10] Google, "WebRTC FAQ," http://www.webrtc.org/faq#TOC-What-is-the-iSAC-audio-codec-.

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