Andrew Hines - Jan Skoglund - Anil C. Kokaram - Naomi Harte *

# MONITORING VOIP SPEECH QUALITY FOR CHOPPED AND CLIPPED SPEECH

*Real-time monitoring of speech quality for VoIP calls is a significant challenge. This paper presents early work on a no-reference objective model for quantifying perceived speech quality in VoIP. The overall approach uses a modular design that will be able to help pinpoint the reason for degradations as well as quantifying their impact on speech quality. The model is being designed to work with narrowband and wideband signals. This initial work is focused on rating amplitude clipped or chopped speech, which are common problems in VoIP. A model sensitive to each of these degradations is presented and then tested with both synthetic and real examples of chopped and clipped speech. The results were compared with predicted MOS outputs from four objective speech quality models: ViSQOL, PESQ, POLQA and P.563. The model output showed consistent relationships between this model's clip and chop detection modules and the quality predictions from the other objective speech quality models. Further work is planned to widen the range of degradation types captured by the model, such as non-stationary background noise and speaker echo. While other components (e.g. a voice activity detector) would be necessary to deploy the model for stand-alone VoIP monitoring, the results show good potential for using the model in a realtime monitoring tool.*

***Keywords:*** *Speech Quality Model, Clip, Chop, VoIP.*

## 1. Introduction

As digital communication has become more pervasive, the variety of channels for human speech communication has grown. Where narrowband telephony dominated, the range of channels has expanded to include multimedia conferencing such as Google Hangouts, Skype and other voice over internet protocol (VoIP) services. Realtime assessment of the Quality of Experience (QoE) for users of these systems is a challenge as the channel has become more complex and the points of failure have expanded. Traditionally, QoE for voice communication systems is assessed in terms of speech quality. Subjective listener tests establish a mean opinion score (MOS) on a five point scale by evaluating speech samples in laboratory conditions. Aside from being time consuming and expensive, these tests are not suitable for realtime monitoring of systems.

The development of objective models that seek to emulate listener tests and predict MOS scores is an active topic of research and has resulted in a number of industry standards. Models can be categorised by application, i.e. planning, optimisation, monitoring and maintenance [1]. Full reference objective models, such as PESQ [2] and POLQA [3], predict speech quality by comparing a reference speech signal to a received signal and quantifying the difference between them. Such models can be applied to system optimisation but are constrained by the requirement to have access to the original signal, which is not always practical for realtime monitoring systems. In these scenarios, no-reference (NR) models, such as P.563 [4], LCQA [5] or ANIQUE+ [6] are more appropriate. They are sometimes referred to as single ended, or non-intrusive models, as they attempt to quantify the quality based only on evaluating the received speech signal without access to a clean reference. This restriction makes NR model design more difficult, and NR models tend to have inferior performance accuracy, when compared to full reference models [7].

This work presents the early stage development of an NR speech quality model for VoIP applications based on a modular architecture. The model will contain modules that are designed to detect and estimate the amount of degradation caused by specific issues. Ultimately the individual modules will be combined to produce an aggregate objective speech quality prediction score. The novelty of this approach over other NR models [4, 5 and 6] is that each module provides a unidimensional quality index feeding into the overall metric but can also provide diagnostic information about the cause of the degradation for narrowband

* [1,2]Andrew Hines, [3]Jan Skoglund, [3]Anil C. Kokaram, [2]Naomi Harte
[1]Dublin Institute of Technology, Ireland
[2]Trinity College Dublin, Ireland
[3]Google, Inc., Mountain View, CA, USA
E-mail: andrew.hines@dit.ie

or wideband speech. This could allow realtime remedial action to be taken to improve the overall quality of experience for the users of VoIP systems, through changing parameters such as bandwidth to adjust the quality of experience from a low quality wideband speech scenario to an alternative high quality narrowband speech scenario.

The modules proposed in this paper, as part of an overall system, are designed to work with narrowband and wideband signals. The two modules are a model sensitive to amplitude clipping and another for choppy speech. These are two common problems in VoIP. Section 2 describes these degradations and their causes. Section 3 describes the models and an experimental evaluation is outlined in section 4. Results are presented for both synthesised and real degradations. Section 5 discusses the results and compares them with the predictions of other objective metrics. The paper concludes with a description of the next stages in the overall model development.

## 2. Background

### 2.1. Amplitude Clipped Speech

Amplitude clipping is a form of distortion that limits peak amplitudes to a maximum threshold. This can be caused by analogue amplifiers where the amplification power exceeds the capabilities of the hardware. Amplitude clipping can also be caused by digital representation constraints when a signal is amplified outside the range of the digital system. If the maximum range of the signal cannot be represented using the number of quantising intervals available (number of bits per sample), the signal will be clipped. The main body of literature studying the effect of amplitude clipping on speech quality is in the field of hearing aids. For hearing aids, clipping can be can used to minimise the distortion for high level input signals [8], whereas in VoIP scenarios, clipping is generally an undesirable result of incorrect gain level settings for the speaker's hardware. The term 'clipped' is often used to describe other types of speech quality degradation, such as time clipped (choppy) or temporally clipped (front end clipping, back end clipping of words) but here it will be used to refer exclusively to amplitude clipping.

Clipping has significantly more impact on quality than intelligibility. Experiments by Licklider [9] found that word intelligibility remained over 96% when speech was clipped to 20 dB below the highest peak amplitude. To put this in perspective, the highest clipping level used in this paper was clipped to 16 dB below the highest peak amplitude and while it is fully intelligible, informal listening tests show it was perceived as very poor quality.

Examples of the clipped speech used in testing are shown in Fig. 1. The first example is clearly clipped as there is a clear threshold amplitude cut-off. The second example shows the same speech with narrowband 30 dB SNR pink noise added after clipping. This illustrates how clipping that is still apparent to the listener can be masked in the signal amplitude by other degradations.

### 2.2. Choppy Speech

Choppy speech describes degradation where there are gaps in the speech signal. It manifests itself as syllables appearing to be dropped or delayed. The speech is often described as a stuttering or a staccato. It is sometimes referred to as time clipped speech, or broken voice. It is generally periodic in nature, although the rate of chop and duration of chops can vary depending on the cause and on network parameters.

Choppy speech occurs for a variety of reasons such as CPU overload, low bandwidth, congestion or latency. When frames are missed or packets dropped, segments of the speech are lost. This can occur at any location within speech, but is more noticeable and has a higher impact on perceived quality when it occurs in the middle of a vowel phoneme than during a silence period. Modern speech codecs attempt to deal with some quality issues by employing jitter buffers and packet concealment methods (e.g. [10 and 11]) but do not deal with all network or codec related problems and choppy speech remains a problematic feature of VoIP systems [12].

## 3. Models

### 3.1. Amplitude Clipped Speech Detection Model

The module is a non-intrusive single ended model. It takes a short speech signal as input and bins the signal samples by amplitude into 50 bins. Two additional bins are added with values set to the minimum bin value to allow first and last bins to be evaluated as peaks. The resulting histogram for a clipped signal is illustrated in Fig. 1 where, $h[i]$, is the histogram value of peak index i. The model finds all local maxima peaks in the histogram. Local maxima peaks are constrained to a minimum height of 0.5% of the sum of the histogram and a minimum distance of 5 bins separation from other peaks. As a minimum of three bins are required to identify a peak, this constraint ensures small deviations in local maxia are not treated as new peaks. Next, all peaks are sorted into descending order yielding a set, $P$. Then, beginning with the largest peak, all peaks not separated by 5 or more bins are discarded. First, the centre peak and clipped peaks, illustrated in Fig. 2 are identified. The centre peak, $P_c$, where $P_c = h[c]$. The peak index, c, is found using auto-correlation of $h[i]$, from

$$c = \frac{1}{2} \arg\max_j R_{hh}[j] = \sum_i h[i]h[i-j] \quad (1)$$

The left peak $P_l$ is found as the largest of the peaks to the left of the centre peak $P_c$, located at

$$l = \arg\max_i h[i] \;\; \forall \;\; i < c, i \in \{P\} \tag{2}$$

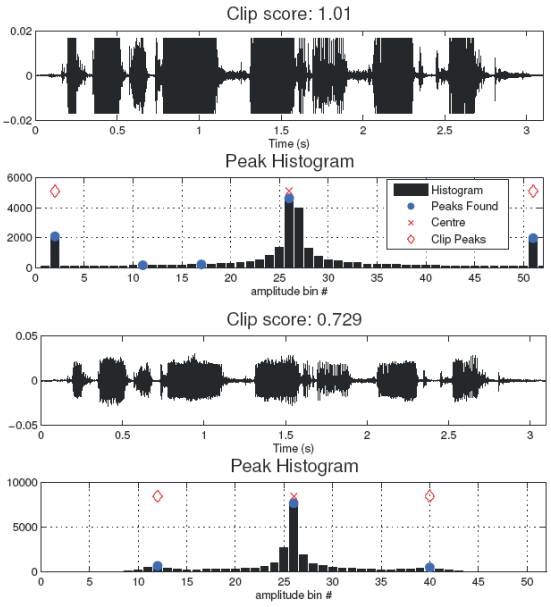


*Fig. 1 Amplitude clipping signal and histogram in the time domain binned across 50 amplitude bins. Above: A signal with clipping visually apparent in the time domain. The histogram highlights the clipping with peaks in the first and last bins. Below: the clipped signal which has been further corrupted with 30 dB SNR narrowband pink noise after clipping. The clipping becomes harder to observe in the signal but the clipping peaks are still visible in the histogram.*

Then the equivalent right peak $P_r$ is the peak closest to the same distance from the centre peak as the left peak, calculated as $h[r]$ where

$$r = \arg\min_i \left| ((c-l)-(i-c)) \right| \;\; \forall \;\; i > c, i \in \{P\} \tag{3}$$

The clip score is calculated as

$$clip = \log_{10}\left[ \frac{\sum_{i=l-1}^{l+1} h[i] + \sum_{i=r-1}^{r+1} h[i]}{\sum_{i=c-1}^{c+1} h[i]} \right] \tag{4}$$

Figure 2 illustrates an example histogram with the maximum peak, $P_c$ and the clip peaks, $P_l$ *and* $P_r$, as solid red bars and other candidate peaks as solid black bars.

### 3.2. Choppy Speech Detection Model

The chop detection model [13] uses a short-term Fourier Transform (STFT) spectrogram of the test signal to measure changes in the gradient of the mean frame power. An example is shown in Fig. 3. The STFT is created using critical bands between 150 and 8,000 Hz for wideband speech or 3,400 Hz for narrowband speech. A 256 sample, 50% overlap Hanning window is used for signals with 16 kHz sampling rate and a 128 sample window for 8 kHz sampling rate to keep frame resolution temporally consistent.

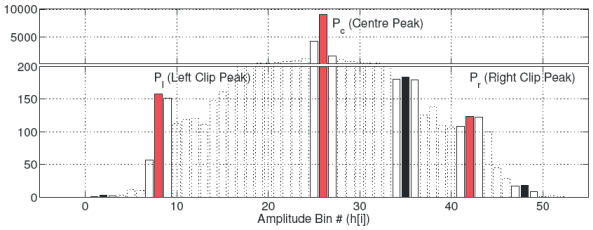


*Fig. 2 Amplitude clipping algorithm. The signal is binned by amplitude into 50 bins. The peaks from the histogram are shown as solid bars. After the centre peak is found using autocorrelation, the left peak is the max peak left of the centre peak. The matching right peak is the peak closest to the same distance from the centre as the left peak. The clipped score is then calculated as a log of the sum of the clip peak bins and their adjacent bins are divided by the sum of the centre peak and adjacent bins.*

A gradient of the mean power per frame is calculated, g[i], as

$$g = \nabla P = \frac{\partial P}{\partial t}. \tag{5}$$

A positive gradient signal, $g_p[i]$ and a negative gradient, $g_n[i]$ can be defined as

$$g_p[i] = \begin{cases} g[i] & \text{if } g[i] > 0 \\ 0 & \text{if } g[i] \le 0 \end{cases}$$
$$g_n[i] = \begin{cases} -g[i] & \text{if } g[i] < 0 \\ 0 & \text{if } g[i] \ge 0 \end{cases} \tag{5}$$

A cross-correlation of $g_p[i]$ and $g_n[i]$ yields the max overlap offset j as

$$\arg\max_j R_{g_n g_p}[j] = \sum_i g_n[i] g_p[i-j]. \tag{6}$$

The $g_p[i]$ and the offset $g_n[i-j]$ are summed as

$$g_c[i] = g_n[i] + g_p[i-j] \tag{7}$$

and a log ratio of the sum of values above a threshold $c_T$ denoted $c_+$, to the sum below the threshold, $c_-$, is taken to estimate the amount of chop in the signal:

$$c_+[i] = \begin{cases} g_c[i] & \text{if } g_c[i] > c_r \\ 0 & \text{if } g_c[i] \le c_r \end{cases}$$
$$c_-[i] = \begin{cases} g_c[i] & \text{if } g_c[i] < c_r \\ 0 & \text{if } g_c[i] \ge c_r \end{cases} \quad (8)$$
$$chop = \log_{10} \frac{\Sigma_i c_+[i]}{\Sigma_i c_-[i]}$$

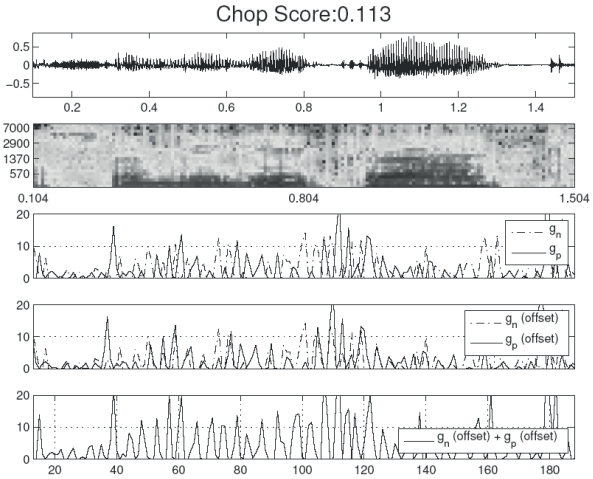


*Fig. 3 Real Chop Example. This example is taken from a real recording where choppy speech occurred as a result of a codec mismatch between transmitter and receiver. The top panes show the signal and signal spectrogram, and the chop is visible as periodic white bands in the higher frequencies of the spectrogram. The gradients, gp and gn are shown in the next pane with the offset versions that have been aligned shown in the forth pane. The bottom pane shows the sum of the offset gradients. This has sharp peaks corresponding to the chop and is used to calculate the chop score, as described in section 3.*

## 4. Model testing

### 4.1. Stimuli

For these experiments a test dataset was created using 30 samples from the IEEE speech corpus [14]. Ten sentences from three speakers, each of approximately 3 seconds in duration were used as source stimuli. A cursory validation with a small number of real clipped and chopped speech samples was also undertaken using wideband recordings of choppy speech caused by a codec mismatch and clipped speech recorded using a laptop microphone.

### 4.2. Model Comparison

The test data was evaluated using 4 other objective speech quality models: ViSQOL, PESQ, POLQA and P.563. ViSQOL is a full reference objective model developed by the authors in prior work [15, 16 and 17]. PESQ [2] is the ITU recommended standard and is still the most commonly used speech quality model although it has be superseded by a newer standard POLQA [3]. P.563 is the ITU standard no-reference model [4].

### 4.3. Amplitude Clipping Test

Each sentence was used to create 20 progressively degraded samples of clipped speech. For each sentence, the peak amplitude was found and the signals were clipped to a factor of the maximum peak amplitude ranging from 0.5 to 0.975 in 0.025 increments. For comparison, this is a range of 13.4 to 0.83 dB re RMS or a clipping threshold 3 dB to 16 dB below the maximum peak.

A second test used the same clipping samples but added narrowband 30 dB SNR pink noise to the signal after clipping. This was done to simulate the realities of amplitude clipping where the signal may be scaled or subjected to additional noise and or channel effects after the clipping occurred. Pink noise was chosen as it has similar spectral qualities to speech. At a 30 dB SNR level, it would not be expected to have a major impact on quality but it will mask the sharp cutoff level of the clipping, as illustrated in the signal plots of Fig. 1.

The 20 sets of stimuli created for the choppy speech detection were also used as input to test the amplitude clipping detection model. These were used to establish a minimum detection threshold boundary and to ensure that the model was only detecting the expected degradation type.

A limited test was carried out with a real recording of clipped data. A foreground speech sample spoken into a microphone over background television speech was recorded. The background speech is not clipped but the foreground speech has moderate to severe clipping. The model was used to evaluate the sentence in 1 second segments and the results are shown in Fig. 4.

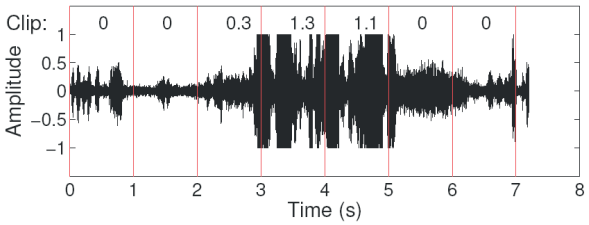


*Fig. 4 Real clipped speech example. A foreground speech sample spoken into a microphone over background television speech was recorded. The background speech is not clipped but the foreground speech (from 2.9-5.1 s) has moderate to severe clipping. The model was used to evaluate the sentence in 1 second segments and the clip scores are marked above each 1 second sample.*

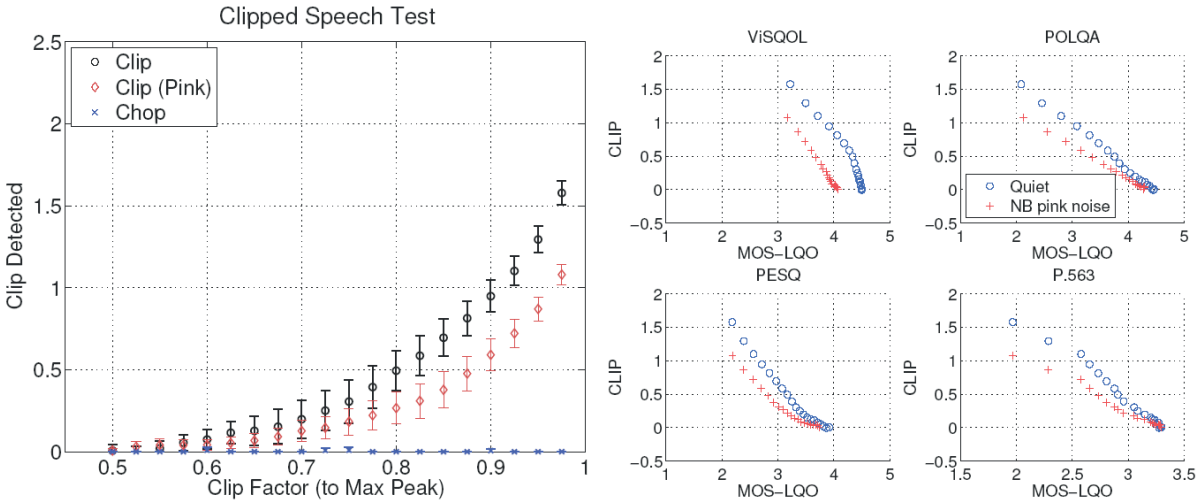


*Fig. 5 Amplitude Clipping Results. Left: Results for clipped speech with the clip level plotted against the clip detected for clipping in quiet and with narrowband pink noise added after clipping. A test is also shown with 20 increasing amounts of chop to investigate the model's detection threshold and sensitivity to other degradation types. Right: Comparison with other objective metrics (both full and NR).*

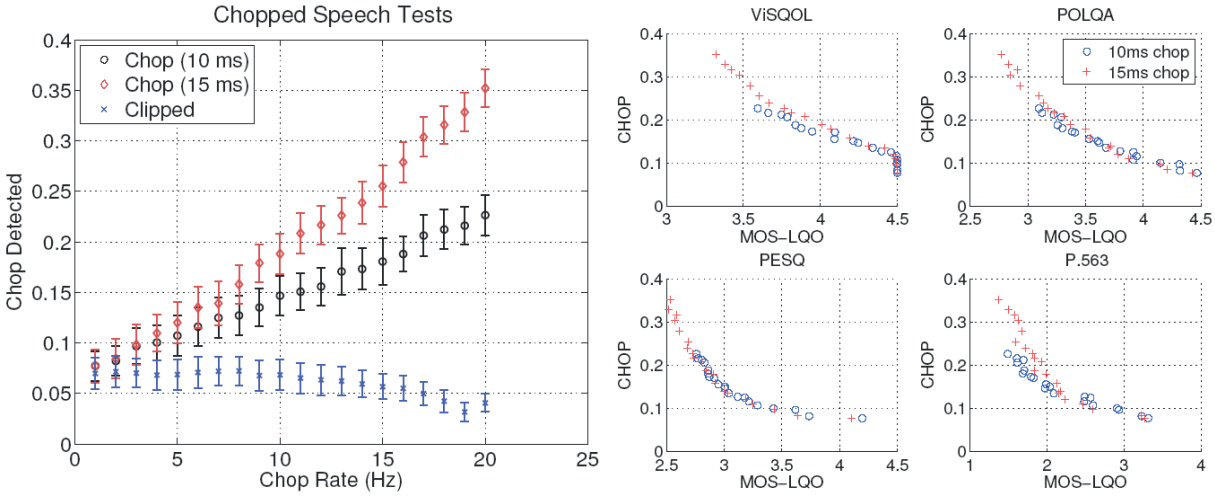


*Fig. 6 Chop Detection Results. Left: Results for choppy speech with the chop rate plotted against the level of chop detected for two chop periods, 10 and 15ms. A test is also shown with 20 increasing amounts of amplitude clipping to investigate the model's detection threshold and sensitivity to other degradation types. Right: Comparison with other objective metrics.*

### 4.4. Choppy Speech Detection Test

Two tests were carried out using chopped speech. Using the 30 source sentences, twenty degraded versions of each sentence were created using two chop frame periods of 10ms and 15ms. This simulated packet loss from 3% to 32% of the signals. The test did not simulate packet loss concealment so the samples for the chopped frames were set to zero.

As with the amplitude clipping test, the chop detection model was cross-validated with the clipped stimuli to establish a minimum detection threshold boundary and to ensure that the model was only detecting the expected degradation type.

A limited test was carried out with real choppy data. Wideband speech with a severe amount of chop was tested. The chop in the test was caused by a codec mismatch between the sender and receiver systems. A segment of the test signal is presented in Fig. 3.

## 5. Results and discussion

### 5.1. Amplitude Clipped Speech

Figure 5 presents the results for the amplitude clipping tests in quiet and pink noise. The level of clipping increases from left to right on the x-axis and the y-axis shows the model output score. The trends in both the quiet and additive pink noise show clipping begins to be detected at clip level of around 0.55 times peak amplitude. This is a 12 dB peak-to-average ratio which was reported by Kates (1994) to be the level at which clipped speech is indistinguishable from unclipped speech.

The chopped data points are shown on the same x-axis for simplicity but are not clipped in any way and represent 20 levels of progressive chop. They are reported here to highlight that the model is not sensitive to temporal or frequency degradations.

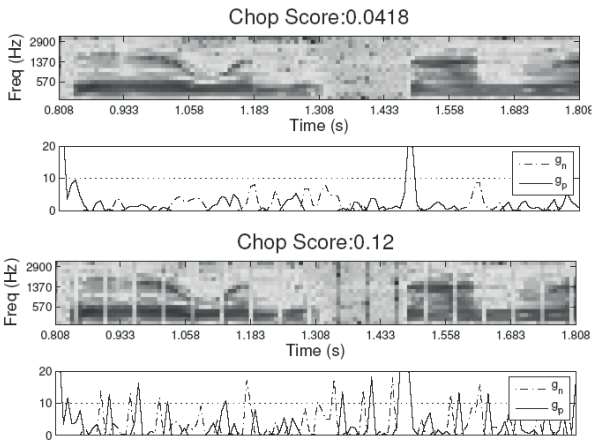


*Fig. 7 Chop Example: Above: Clean speech signal with gradients gp and gn plotted below. The gradients detect the gradient in the speech with a large gradient change visible at approximately 1.5 s. Below: The same speech with chop added. The chop is visible in the spectrogram and visible in the $g_p$ and $g_n$ plot used to calculate the chop score.*

Although the trends are similar, the range of the clip scores for the quiet and pink noise are different. This is due to the relationship between the scale and the count in the histogram bins. The difference in height between the sharp peaks seen in the quiet histogram versus the spread of peaks in the noisy histogram can been seen in Fig. 1. The use of the additional bins either side of the clip peaks and centre peaks in the ratio calculation (4) reduced the overall difference between the model estimate for a given clipping level when measured in quiet or with additive noise.

Figure 5 also presents a comparison between the model output and those of four other objective quality metrics: ViSQOL, PESQ, POLQA and P.563. For reference, the MOS-LQO predictions are presented in Fig. 8. The results for POLQA in Fig. 5 show

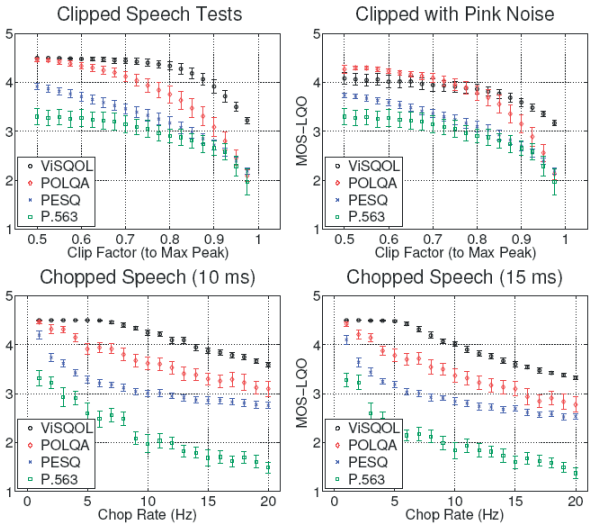


*Fig. 8 Predicted MOS-LQO from objective metrics compared in Fig. 5 for clip tests and Fig. 6 for chop tests. Results have shown mean results over 30 sentences. Error bars are 95% confidence intervals.*

a linear relationship between the clip scores and the objective metrics across the full range of tests while ViSQOL, PESQ and P.563 exhibit a variety of different sensitivities for the tests with low amounts of clipping, leading to nonlinear tails in the plots. It is worth noting in Fig. 8 that the addition of pink noise to the clipped signal had little effect on the POLQA results for peak clip factors from 0.50–0.6 whereas PESQ and ViSQOL results dropped by over 0.5 with the pink noise added.

### 5.2. Choppy Speech

Figure 6 presents the results for the chopped speech. The chop rate increases from left to right on the x-axis and the y-axis shows the model output score. The results for the amplitude clipped speech are shown on the same x-axis for simplicity but are not chopped in any way and represent 20 levels of progressive amplitude clipping. They highlight that there is a lower threshold to the chop detection. Fig. 7 shows the same speech sample with and without chop. The periodic chop is clearly visible as vertical bands across the spectrogram and in the peaks of the negative and positive gradients, gp and gn, used by the model to estimate the signal chop level. In addition to detecting chop, the natural gradients of speech are captured by the model. The natural gradient at 1.5 seconds is very apparent in Fig. 7. These speech features are responsible for the low threshold boundary of the chop detection model. The trend for both chop frame periods show chopping being detected above the threshold from a chop rate of 2 Hz. Chop at low rates are common in practice so preliminary tests (not presented here) were carried out with longer duration speech samples. They showed that better

separation between results for chop and naturally occurring gradient changes is possible. This constraint would present practical implementation challenges in a realtime monitoring implementation but should not be insurmountable.

Fig. 8 also presents a comparison between the model output and those of four other objective quality metrics: ViSQOL, PESQ, POLQA and P.563. Unlike the results for the clipping model, the chop model does not have a linear relationship with the objective model results. However, the curve is quite consistent across the different model comparisons, meaning a simple quadratic regression fitting from the chop model score to a MOS prediction may be sufficient. The 10 ms and 15 ms chop periods follow linear trends in Fig. 6 but with different slopes. When they are plotted against the objective metrics there is an overlap in the results follow the same curve. This represents a strong relationship between the chop models score and the estimated perceived quality from the objective metrics.

The real chop example tested showed that chop is detected even if the chop value is not zero and the chop frame is shorter than 10ms, as was the case in the simulated chop tests.

## 6. Conclusions and future work

The clip and chop measurement models for speech quality presented in this paper show promising early results and compares favourably to the other no-reference objective speech quality model. The degradation types detected are common problems for VoIP and the algorithms used are relatively low in computational complexity. These factors, combined with their applicability to both narrowband or wideband speech, mean they could be useful in applications other than full speech quality models, for example as stand-alone VoIP monitoring tools. To use the model in a realtime system, other components would be necessary. For example, the chop or clip detection will not give accurate results if the speech contains large segments of silence. This could easily be addressed with voice activity detection prior to chop and clip detection.

The models presented are still in the early stages of development. They require testing with a broader test set including a wide variety of real rather than generated degradations. Further testing with a range of wideband stimuli is also required. MOS tests on the existing data would also be beneficial as the full reference metrics disagree significantly on their MOS-LQO predictions for both the clipped and choppy speech. The correlation with quality predictions from POLQA was stronger than with the other objective models. This is seen as a positive pointer for the performance against subjective listener test results as POLQA reports better accuracy than PESQ and and has become the new benchmark standard.

## References


[1] MOLLER, S., CHAN, W-Y., COTE, N., FALK, T. H., RAAKE, A., WALTERMANN, M.: Speech Quality Estimation: Models and Trends, *Signal Processing Magazine, IEEE*, vol. 28, No. 6, pp. 18-28, 2011.

[2] 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, CH, ITU-T Rec. P.862, 2001.

[3] ITU: *Perceptual Objective Listening Quality Assessment*, Int. Telecomm. Union, Geneva, CH, ITU-T Rec. P.863, 2011.

[4] ITU: S*ingle-ended Method for Objective Speech Quality Assessment in Narrow-band telephony Applications,* Int. Telecomm. Union, Geneva, CH, ITU-T Rec. P.563, 2004.

[5] GRANCHAROV, V., ZHAO, D. Y., LINDBLOM, J., KLEIJN, W. B.: Low-complexity, Nonintrusive Speech Quality Assessment, *IEEE Audio, Speech, Language Process.,* vol. 14, No. 6, pp. 1948-1956, 2006.

[6] ANSI ATIS: 0100005-2006: *Auditory Non-intrusive Quality Estimation Plus (ANIQUE+): Perceptual Model for Non-intrusive Estimation of Narrowband Speech Quality,* American National Standards Institute, 2006.

[7] FALK, T. H., CHAN, W-Y.: Single-ended Speech Quality Measurement Using Achine Learning methods, *IEEE Audio, Speech, Language Process.,* vol. 14, No. 6, pp. 1935-1947, 2006.

[8] STELMACHOWICZ, P. G., LEWIS, D. E., HOOVER, B., KEEFE, D H.: Subjective Effects of Peak Clipping and Compression Limiting in Normal and Hearing-impaired Children and Adults, *J. Acoust Soc Am*, vol. 105, pp. 412, 1999.

[9] LICKLIDER, J. C.: Effects of Amplitude Distortion Upon the Intelligibility of Speech, *J. Acoust Soc Am,* vol. 18, No. 2, pp. 429-434, 1946.

[10] BENESTY, J., SONDHI, M. M., HUANG, Y. A. : *Springer Handbook of Speech Processing*, Springer, 2007.

[11] GOOGLE: *WebRTC NetEQ Overview*, http://www.webrtc.org/reference/architecture #TOC-NetEQ-for-Voice.

[12] RAAKE, A.: *Speech Quality of VoIP - Assessment and Prediction*, Wiley, 2006.

[13] HINES, A., SKOGLUND, J., HARTE, N., KOKARAM, A. C.: *Detection of Chopped Speech*, Patent, US20150199979 A1, 07 2015.

[14] IEEE: *IEEE Recommended Practice for Speech Quality Measurements, Audio and Electroacoustics*, IEEE Transactions on, vol. 17, No. 3, pp. 225-246, Sep 1969.

[15] HINES, A., SKOGLUND, J., KOKARAM, A. C., HARTE, N.: VISQOL: An Objective Speech Quality Model, *EURASIP J. on Audio, Speech, and Music Processing,* vol. 2015:13, May 2015.

[16] HINES, A, SKOGLUND, J, KOKARAM, A. C., HARTE, N.: *VISQOL: The Virtual Speech Quality Objective Listener,* IWAENC, 2012.

[17] HINES, A, SKOGLUND, J, KOKARAM, A. C., HARTE, N.: *Robustness of Speech Quality Metrics to Background Noise and Network Degradations: Comparing ViSQOL, PESQ and POLQA,* Acoustics, Speech, and Signal Processing, IEEE Intern. Conference on (ICASSP '13), 2013.