Incoherence of the Signal-Based Passive Speech Quality Estimation Models

Adil Raja*, Colin Flanagan

Abstract—Over the recent past there has been a significant improvement in devising perceptual models for passively estimating the speech offered by telecommunication networks. Telephony networks, on the other hand, are being adapted to support voice over IP. To date, listening quality of voice over IP remains dominated by packet loss. In this paper we claim that passive signal-based speech quality estimation models are inherently incapable of capturing a packet loss event. We support our claim in the light of the underlying DSP-based techniques.

Index Terms—speech quality, genetic algorithms, genetic programming, symbolic regression.

I. INTRODUCTION

Speech quality estimation is vital to the evaluation of quality of service offered by a telecommunication network. Traditionally, speech quality is estimated using subjective tests. In subjective tests, the quality of a speech signal under test is evaluated by a group of human listeners who assign an opinion score on an integral scale ranging between 1 (bad) to 5 (excellent). The average of these scores, termed the Mean Opinion Score (MOS), is considered as the ultimate determinant of the speech quality [1]. Subjective tests are, however, time consuming and expensive. To make up for these limitations, there has been a growing interest in devising software based objective assessment models. There are, further, two kinds of objective assessment models, namely, intrusive and non-intrusive. Intrusive models evaluate the quality of a distorted speech signal in presence of a corresponding reference signal. The current ITU-T recommendation P.862 (PESQ) is an example of such an approach. Non-intrusive models, on the other hand, do not enjoy this privilege and base their results, instead, solely on the estimated features of the signal under test. For this reason, the results of the latter type of models are generally considered inferior to the those of the former type.

Non-intrusive models can further be classified either as signal-based models or the parametric ones. The signal-based models process the distorted speech signal by various techniques such as modeling the human speech production system, auditory signal processing and/or other waveform processing.
techniques. Various features pertaining to the signal under test are extracted and their values are mapped on to human perception of speech quality. An example if such a model is the current, state-of-the-art, ITU-T Recommendation P.563 (PSEAM) for single-ended estimation of speech quality [2].

Parametric models do not process the speech signal rather base their results on various properties relevant to the telecommunications network. In the case of Voice over IP (VoIP), these may be transport layer metrics such as packet loss, jitter and end-to-end delay of a call. Such models are considered to be suitable for real-time evaluation of call quality. However, they may fail to present a realistic view of the quality of service being offered by a communications network where such distortions are prevalent that are not catered for by the parametric model. An example of a parametric model is the ITU-T G.107, commonly referred to as the E-model [3].

In this paper it is shown that the signal-based non-intrusive models are inherently incapable of detecting a packet loss event. Thus, this also leads to their inability to estimate the speech quality in an event where packet loss is the dominant cause of signal distortion. In our approach we performed an experiments where a large number of speech samples are distorted under varying degrees of Bernoulli and burst losses. The resulting, degraded, speech samples are processed with both PESQ and ITU-T P.563 (or P.563 for brevity) to search for any correlation between the results of the two algorithms. Parameters of pertaining to packet loss are also analyzed for their utility and efficiency in detecting loss events. We also argue about inefficacy of alternative approaches of signal-based quality estimation and present reasons.

Rest of the paper is organized as follows. In section II, we discuss recent developments in non-intrusive speech quality estimation research with an emphasis on signal-based models. Section III discusses the various reasons due to which signal-based models fail in discovering packet loss events. The experimental setup and the consequent results are described in section IV. Some speculations pertaining to other approaches advanced are listed in Section V. Section VI presents the conclusions.

II. NON-INTRUSIVE QUALITY EVALUATION

Numerous non-intrusive speech quality assessment schemes have been proposed in the past. In [4] Jin and Kubichek proposed to form a codebook of Perceptual Linear Prediction (PLP) feature vectors of clean speech using vector quantization (VQ). The ultimate speech quality is a function of the mismatch between the feature vector of the distorted speech signal and its most similar counterpart in the reference codebook. In [5] an error backpropagation Multi-Layer Perceptron (MLP) was trained to map the psycho-acoustically relevant features of speech on to human assessment of speech quality. Similarly,

In May 2004 ITU set a new standard for non-intrusive evaluation of speech quality, referred to as ITU-T Recommendation P.563 [2]. The algorithm aims at estimating the subjective quality of narrow-band speech signals that might be distorted due to background noise, filtering, frame loss/erasures, faulty communications channels and low bit-rate codecs.

A contemporary to P.563 is the ANIQUE model [8] which was marginally outperformed by the former in a competition held by ITU-T in 2004 to select a non-intrusive speech quality estimation model.

Recently, Falk and Chan in [9] and Grancharov et al. in [10] have proposed alternative speech quality estimation models to refute the performance claims of P.563. Albeit their disparate speech feature specification and extraction methodologies, mapping between speech features and subjective quality is derived using Gaussian Mixture Models (GMMs). As a crux, signal-based models are mainly based on approaches that depict the human sound production system or the peripheral or central auditory signal processing system.

While these models have claimed to propose promising results for various distortion types including frame loss and frame erasures, we claim that signal-based models are inherently inept at detecting the presence of a frame loss/erasure event.

P.563 may also supposedly estimate the frame repetition. Frame repetition occurs when a codec’s packet loss concealment (PLC) algorithm repeats the contents of a previously successfully received frames in the event of a packet/frame loss or erasure. The repeated frames are intended to camouflage the effect of lost frames on listeners’ perception of speech quality. The various commendable performance details of the algorithm are discussed in detail in [11].

III. REASONS OF INEPTNESS

In the most primitive case, a packet/frame loss/erasure event marks an absentee of that portion of the speech from the waveform eventually received on the remote end-point. The underlying speech codec may as well fill this gap with comfort noise. The overall effect of loss may lead to a distortion that is audible to human listener. However, this may not be discernible to a signal-based non-intrusive model as it may fail to ascertain whether the portion of waveform having no speech activity, or having comfort noise, represents silence of the talker or a loss event. P.563 attempts to alleviate this problem by extracting various features based on time domain characteristics of the speech signal. More specifically, it performs temporal clipping detection whereby it attempts to discern abnormal muting or interruption of the speech.
Another typical case is where a codec is equipped with a Packet Loss Concealment (PLC) algorithm and the PLC algorithm, in a meaningful sense, repeats the contents of the previous frame in the event of a packet loss. The repeated frames are intended to camouflage the effect of lost frames on listeners’ perception of speech quality. The PLC mechanism of G.729 may be considered as a case study. In the event of a frame loss/erasure, the Line Spectral Frequencies (LSF) for the lost frame are repeated from the previous frame. The adaptive and fixed code-book gains are also taken from the previous frame. The excitation is taken from the previous frame while the excitation component (fixed or adaptive) to be taken depends on whether the previous frame was classified as voiced or unvoiced. Such a scheme is deemed successful in concealing the adverse effect of one or two successively lost packets. However, when more packets are lost consecutively, as in bursty traffic conditions, the repetition of previous frame contents again surmounts to listeners annoyance. To counteract this problem the PLC algorithm of G.729 codec gradually decreases the fixed and adaptive code-book gains while concealing successive losses. Clearly, the motto of the codec is to become silent, in the face of continued packet loss, as opposed to persistent babbling. This phenomenon is pictorially depicted in Figure 1. The 1.66 s long waveform of the phoneme [a] (as in Lahore uttered by one of the authors) was subjected to frame loss using ITU-T G.729 speech codec. First 20 frames of the speech segment were decoded successfully, whereas from there onwards, successive frames were lost until the end of the segment and the codec’s PLC algorithm was utilized to reconstruct lost speech. The resulting signal’s magnitude, as shown in Figure 1(b), started decaying from 0.2 s and died completely before the end of the first second.

Fig. 1. Hi there

P.563 attempts to detect the frame repeats in the speech signal by hoping to exploit the high cross-correlation of the repeated frames. According to the reference implementation of the implementation two adjacent 32 ms frames, with 4 ms overlap, of the signal are processed at a time. After fulfilling rel-
evant criteria, if the cross-correlation coefficient of the signal exceeds a certain value then the value of the FrameRepeats parameter is incremented by one. Similarly, the value of the FrameRepeatsTotEnergy is incremented by the current frame’s Root Mean Squared Energy (RMSE).

However, given that the codecs may gradually decrease their excitation gains in the countenance of frame loss, two problems may arise. First, with the diminishing gain, the waveform of the repeated frames start to vanish. This may hamper the detection of correlation. Secondly, accumulating the RMSE of the frames as a distortion indicator seems anomalous too. As, to a certain extent, the annoyance created by a burst packet loss event increases with its length. Whereas, time diminishing RMSE of the portion of the signal where a burst loss event has occurred tends to contradict with the impact of the distortion. In the next section we strengthen our hypothesis in the light of experimental details.

IV. EXPERIMENT AND RESULTS

An initial experiment was performed to evaluate the performance of the P.563 algorithm in packet loss conditions. A total of 1440 distorted speech samples were created by subjecting 240 clean speech samples to varying levels of bursty packet loss. Speech data was collected from Nortel networks. Each speech sample had a duration of 10–12 sec with speech activity ranging between 70–80%. G.729 was used as the preferred codec in this work. Gilbert Elliot model was used to emulate bursty loss conditions [12]. Packetization interval (PI) was also varied between 10-60 ms, which has an effect similar to that of bursty losses. The resulting speech samples were evaluated using both P.563 and the PESQ algorithms to look for mutual correlation between results. A value of 0.2570 for the Pearson’s product moment correlation coefficient (σ) was observed that depicts no correlation between the two algorithms. Figure 2 also delineates a clear lack of correlation between the results of two algorithms.

![Figure 2. Correlation between ITU-T P.563 and the PESQ algorithm](image)

In order to further scavenge for supporting evidence another experiment was performed. Two more codecs were included in this apart from G.729. These include G.723.1 and AMR-NB. The former was used in its 6.3 kbps mode and the latter in its 7.4 and 12.2 kbps modes. Both of these codecs are also equipped with PLC algorithms. Packet loss was simulated for different values of mean loss rate (mlr) and conditional loss probability (clp), which corresponds to the level of burstiness. The choice of
A model for evaluating multimedia transmission performance over an IP network. Twelve different values for \textit{mlr} were chosen; 0, 2.5, 5, 7.5, 10, 12.5, 15, 20, 25, 30, 35 and 40%. The peak loss-rate (i.e. 40%) was kept an order of magnitude higher than that specified for an unmanaged network in ITU-T G.1050 (i.e. 20%). For each value of \textit{mlr}, conditional loss probability (\textit{clp}) was set to 0.1, 0.5, 0.6, 0.7, and 0.8%. \textit{clp} is used to simulate bursty packet losses in a meaningful sense using the Gilbert model, with a larger value of \textit{clp} meaning a higher degree of burstiness. Moreover, \textit{PI} was also varied between 10–60 ms. A total of 3360 speech samples were prepared for various combinations of afore-mentioned values of parameters. The speech samples were evaluated with the P.563 and PESQ algorithms. The \textit{MOS-LQO} \textsubscript{P563} and alongwith values of cogent metrics of P.563 were calculated. These include \textit{FrameRepeats} and \textit{FrameRepeatsTotEnergy} as discussed in section III. \textit{MOS-LQO} \textsubscript{PESQ} evaluations were based on the mapping function defined by ITU-T Recommendation P.862.1 [13]. It is mentioned for purpose of clarity that the term \textit{MOS-LQO} is an acronym for \textit{Mean Opinion Score–Listening Quality Objective} and the various subscripts are used to identify the objective quality estimation models used. This terminology is based on [14].

Once again a low value of correlation coefficient (\(\sigma=0.5983\)) was obtained between \textit{MOS-LQO} \textsubscript{P563} and \textit{MOS-LQO} \textsubscript{PESQ}.

\textbf{A. Efficacy of P.563 parameters}

Table I lists very low, negative, values of \(\sigma\) between the P.563 features, namely \textit{FrameRepeats} and \textit{FrameRepeatsTotEnergy}, and the \textit{MOS-LQO}.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Stats & \textit{MOS-LQO} \textsubscript{P563} & \textit{MOS-LQO} \textsubscript{PESQ} \\
\hline
FrameRepeats & -0.1699 & -0.0988 \\
FrameRepeatsTotEnergy & -0.0599 & -0.0764 \\
\hline
\end{tabular}
\caption{Correlation between P.563 parameters and MOS-LQO}
\end{table}

To further estimate the efficacy of P.563 parameters, namely, \textit{FrameRepeats} and \textit{FrameRepeatsTotEnergy}, were regressed on to \textit{MOS-LQO} \textsubscript{PESQ} and \textit{MOS-LQO} \textsubscript{P563}. While the erstwhile was performed to estimate the effectiveness of the afore-mentioned parameters in general, the second regression was conducted to analyse the influence of these parameters in forming \textit{MOS-LQO} \textsubscript{P563}. Given that P.563 assumes a linear relationship between its parameters and its eventual speech quality estimate, multiple linear regression was employed to find a befitting mapping. The consequent regression equations in turn also had poor correlation (\(\sigma=0.1727\) and 0.1158) with \textit{MOS-LQO} \textsubscript{PESQ} and \textit{MOS-LQO} \textsubscript{P563} respectively.

Figure 3 clarifies the cornucopia of data on average number of \textit{FrameRepeats} vs \textit{mlr} and \textit{clp} using
a subset of overall test configurations. It can be seen that values of *FrameRepeats* parameter do not have any correlation with *mlr* or *clp*. Moreover, it is also notable that values of *FrameRepeats* parameter at 0% *mlr* are comparable to those at other values of *mlr* and that the number of *Frame Repeats* never exceeds 2, even if *mlr* is as high as 40%.

V. Discussion

VI. Conclusion

References


Fig. 3. Bar plots show the average values of \textit{FrameRepeats} parameter for various combinations of \textit{mlr} and \textit{clp} (a) G.729, PI=10 ms, (b) G.723.1, PI=30 ms, (c) AMR-NB 12.2 kbps, PI=60 ms