OBJECTIVE LISTENING QUALITY ASSESSMENT OF SPEECH COMMUNICATION SYSTEMS INTRODUCING CONTINUOUSLY VARYING DELAY (TIME-WARPING): A TIME ALIGNMENT ISSUE

Ludovic Malfait, Dr Phil Gray,
Psytechnics Ltd,
Ipswich, United Kingdom
{ludovic.malfait, phil.gray}@psytechnics.com

Dr Martin J. Reed
Department of Computing and Electronic Systems,
University of Essex, United Kingdom
mjreed@essex.ac.uk

ABSTRACT

Standardized in 2001, ITU-T Rec. P.862 (PESQ) is the current "in-force" ITU-T standard for intrusive listening speech quality assessment and is widely used in the speech communication industry. It has proved to be reliable in the assessment of the quality of signals transmitted through traditional communication networks (landline, mobile and first generation of voice over IP systems). However, with the development of internet telephony, like Skype or MSN, advanced techniques for packet concealment and jitter buffer adaptation maybe used, introducing continuously varying delay and making PESQ unreliable. This paper finds that the alignment resolution required by the next generation PESQ–like algorithms to accurately predict subjective tests that evaluate systems that incorporate time-warping would be ±5ms.

Index Terms — Objective quality assessment, speech communication systems, continuously varying delay, ITU-T P.862, time alignment

1. INTRODUCTION

There are two commonly used ways to evaluate the quality of a speech signal transmitted through a communication network. The first method, known as subjective testing, is to ask people to listen to the speech signal and to express their opinion regarding its quality. The other method, known as objective testing, uses a computer program for analyzing the signal and predicting the quality. Both of these methods are widely applied in industry for assessing most voice communication technologies. There has been good success with objective models, however, with newer VoIP systems and most notably soft clients, there have been problems applying existing objective models due to "time warping" effects. Such degradations cause difficulties with signal synchronization in some objective models.

Subjective quality assessment is the most reliable (benchmark) method for determining the listening quality of speech signal. ITU-T Recommendation P.800 [1] describes three methods for conducting listening quality subjective tests, of which the ACR (Absolute Category Rating) is the most popular. In such a test, conducted in a controlled environment, a group of subjects listen to speech samples. After each presentation of a sample, subjects are asked to give their opinion on the quality of the sample on the following discrete scale: bad (1), poor (2), fair (3), good (4), excellent (5). The average of the votes across subjects for a given sample is then calculated. The resulting average is called Mean Opinion Score (MOS).

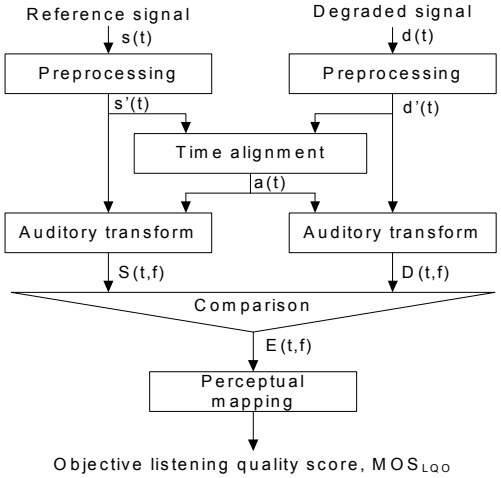

Fig. 1. High-level description of ITU-T Rec. P.862 (PESQ), the currently "in-force" full-reference objective model for listening speech quality.

Objective quality assessment involves algorithms rather than people. It is more cost and time effective than subjective assessment, but it is known to be less reliable. These algorithms are designed for predicting the outcome of subjective tests. There are three main categories of algorithms directly linked to their inputs. The first category is *full-reference* models. These models require the signal injected into the system under test (the reference) in addition to signal to be assessed (the degraded). ITU-T Rec. P.862 [2][3][4], also known as PESQ, is the latest standard in this category. The second category is *no-reference* models. In this case, the algorithm receives the degraded signal only. ITU-T Rec. P.563 [5][6] is the current standard. The third category is *parametric* models. Instead of analyzing the signal itself, these models use the information available regarding the conditions of the network to predict the speech quality. ITU-T Rec. P.564 [7] describes a method for

conformance testing models designed for voice over IP networks.

In this paper, we focus our interest on PESQ, the currently "in-force" full-reference model standardized by the ITU-T. To evaluate the quality of the signal, PESQ compares the reference signal with the degraded signal. Fig. 1 provides an overview the algorithm. After a preprocessing stage that normalizes the signal, the time (mis)-alignment between the reference and the degraded signals is determined and then compensated for. Then the "loudness" of the signals is compared frame-by-frame and the resulting difference is quantified and mapped onto a MOS scale.

Originally designed for assessing narrowband telephony, PESQ was standardized in Feb. 2001. It achieves very high correlation with subjective experiments testing Public Switched Telephone Network (PSTN), mobile and VoIP networks that were available at the time. PESQ was extended to wideband telephony assessment in Nov. 2005 – Wideband PESQ (WB-PESQ). This extension consists only of a modification to the preprocessing filter and auditory transform. Since the standardization of PESQ, Internet telephony has become very popular and advanced techniques for improving the transmission of speech signals on VoIP systems have been developed. Packet loss concealment and jitter buffer adaptation are increasingly performed by extending or shortening speech events (time warping) to compensate for transmission errors. Fig. 2 shows an actual example of a speech signal that has been extended to compensate for an error. While time-warping provides very good performance in terms of speech quality compared to traditional muting or synthesizing techniques, it makes PESQ unable to assess quality accurately. This is because time-warping has the effect of introducing quasi-continuous variable delay between the reference and the degraded signal, making the time alignment much harder to achieve precisely. For an objective model such as PESQ, a few misaligned frames could cause a large error in its output. This performance characteristic has been found in real-world systems. Global IP Solutions exposed problems using PESQ in conjunction with certain VoIP technologies showing its inaccuracy in presence of time-warping [8]. With the increasing use of VoIP by major operators, and the need of operators, customers, and regulators to reliably test these systems, it is imperative that models like PESQ are re-engineered to provide acceptable performance across this new technology. This paper addresses the first step in this process through identification of the scale of the time-alignment problem and quantifying the accuracy required in the time-alignment of a new PESQ like model.

The methodology of this work required a large subjective experiment to be conducted that evaluated several VoIP systems, most of them introducing time-warping. The output of PESQ algorithm was compared with this subjective test, altering its time alignment progressively between a high value of misalignment to lower values until it was (near) perfectly aligned. This process determines the necessary time-alignment accuracy to reach an acceptable performance in the presence of time-warping (or other synchronization distortions) as is required for the next generation of objective assessment tools.

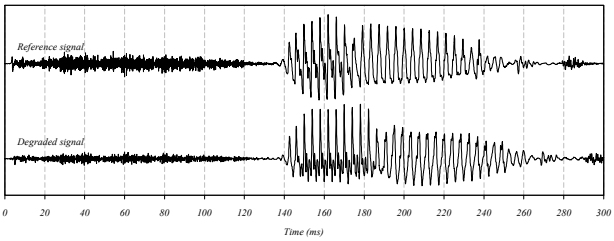

Fig. 2. Example of time-warped section. Both signals are synchronized until the 150th ms, then the degraded signal is stretched, the next pitch cycle being repeated 3 times, introducing a 15ms delay for the rest of the signal.

2. THE SUBJECTIVE EXPERIMENT

The following subjective experiment was conducted in Psytechnics subjective test facilities, a controlled acoustic environment fitted with professional audio equipment. The listening quality of five wideband software VoIP clients (AOL, Google Talk, Skype, Windows Live Messenger, Yahoo Messenger) were evaluated in the presence of various network conditions using the ITU-T Rec. P.830 method [9] on an ACR scale. 24 naïve subjects, divided into 4 listening groups, scored a total of 640 degraded samples altogether, grouped into 40 different network conditions. The conditions were composed of various packet loss and network delay settings through which material was passed for each of the clients under test. In addition to the effect of time-warping in the soft clients, some reference conditions, involving codec distortion only, were included in the test. It should be noted that at the time the experiment was conducted, only one of the five VoIP clients was not using time-warping techniques.

3. WB-PESQ OBJECTIVE MEASUREMENT

The subjective test immediately highlights the problem encountered with PESQ. Fig. 3a presents the output of WB-PESQ against the subjective test showing per sample (termed per file on the graphs) and per condition scores. In such a representation, the points representing the results from an objective model that perfectly predicts the subjective test would form a straight line along $y=x$. On this scatter plot, it is clear that PESQ is far from perfect, achieving a poor correlation of 0.630 per file and 0.696 per condition. Nevertheless, it can be noted that several points are almost aligned on the diagonal. These points correspond to the reference conditions and the VoIP client that does not introduce time-warping. For these conditions, PESQ aligns the reference and degraded signals properly and therefore assesses them correctly. The prediction for the rest of the samples is systematically low. This is due to misalignment

of the frames. As PESQ compares auditory frames one by one, any misalignment would lead to the comparison of unrelated frames and therefore generate unrepresentative errors, lowering the predicted MOS value.

PESQ time alignment has to be improved in order to predict the outcome of this subjective experiment, but by how much? Some studies were undertaken during the development of PESQ; Rix et al. reported that scores from PSQM, the predecessor of PESQ, could drop by a fifth of its output range when frames are misaligned by a constant 10ms offset [10]. The experiment conducted for the investigation in this paper differed as the delay effects were compounded by the variation introduced by the time warping. It is worthwhile considering the delay/signal conditions that might, hypothetically, cause a problem. The size of each PESQ auditory frame is 32ms and a Hanning window is applied before computing the loudness estimation. As a result, the amount of signal processed for each frame is about half the size of the frame, e.g. 16ms. For a non periodic signal, a misalignment of more than ±8ms might lead to a comparison of unrelated signals. However, speech signals are pseudo-periodic over short periods when voiced events occur. A misalignment of more than 8ms might not be an issue in this case. On the other hand, a misalignment of even 1ms could be critical during a transient period. The next section considers the methodology applied to determine the accuracy actually required.

4. METHODOLOGY

In order to rectify PESQ time alignment, a near perfect delay profile was needed for each of the samples in the subjective test. This was generated by means of a tool, developed by the authors, that first pre-aligns signals and then allows manual adjustments to check that the delay profile was correctly identified. It is not trivial to align degraded signals precisely (to the sample) as waveforms can be too different or unrelated during degradations. Nevertheless, the authors are highly confident that the generated delay profiles are accurate to the millisecond, which has proved sufficient for this investigation.

Having both the original PESQ time alignment and the near perfect time alignment information it is then possible to artificially manipulate the PESQ time alignment between the poor PESQ algorithm alignment and the near perfect alignment. This was performed by capping the maximum PESQ alignment error in progressively smaller levels; the results show capped alignment in ms units of: none (normal PESQ), 30, 20, 10, 5, and the near perfect (less than 1ms error). In other words, for the 30ms capped comparison, all alignment errors worse than 30ms were adjusted to 30ms errors, before PESQ produced its quality assessment. Comparing the outputs from the externally modified PESQ algorithm with the results from subjective tests provides the key objective of this paper.

5. RESULTS

Fig. 3 shows the performance of the PESQ algorithm for the different maximum misalignment errors (capping values). The distribution of the amount of corrected frames, calculated on speech sections only, shows that more frames are misaligned when the subjective score is low. This can be explained by the fact that when network conditions are poor, packet loss and network delay increase and packet concealment and jitter buffer adaptation are more frequent.

Error analysis on a frame-by-frame basis shows that, for this subjective test, 6.01% of the frames containing speech information were misaligned by more than 30ms, 10.30% by more than 20ms, 14.69% by more than 10ms and 22.40% by more than 5ms.

During the competitive standardization activity for ITU-T Rec. P.862, a requirement for the proponent models was to achieve a minimum correlation of 0.90 per condition and thus this seems a reasonable target threshold for this investigation. In our experiment, 0.93 is reached if the misalignment never exceeds 10ms. However, Fig. 3e shows that the model could perform better still if the alignment error is less than 5ms as the correlation raises from 0.936 to 0.973. As the performance improvement from 5ms to 1ms maximum misalignment is not significant (0.973 to 0.974), the need for a perfect time alignment is not critical and ±5ms time alignment accuracy proved adequate for assessing time-warped signals correctly.

Even on the graph showing the most accurate time alignment, some outliers can be identified. They are due to the fact that the impact of time-warping on the perception of the quality, even if correct alignment within the model is achieved, is not taken into account. This investigation provided WB-PESQ with a corrected time alignment but the perceptual effect of compressing or extending words (when subjects found it disturbing) was not part of this study. It is therefore anticipated that further performance improvement can be achieved by considering this last point.

6. CONCLUSION

In this paper, the authors presented results from WB-PESQ, the full-reference listening quality assessment model standardized by the ITU-T. This wideband model was applied to speech signals transmitted through modern communications networks introducing continuous variable delay. The output of the wideband model was compared to subjective scores and showed poor performance. We showed in this paper that time-alignment is the main cause of the problem.

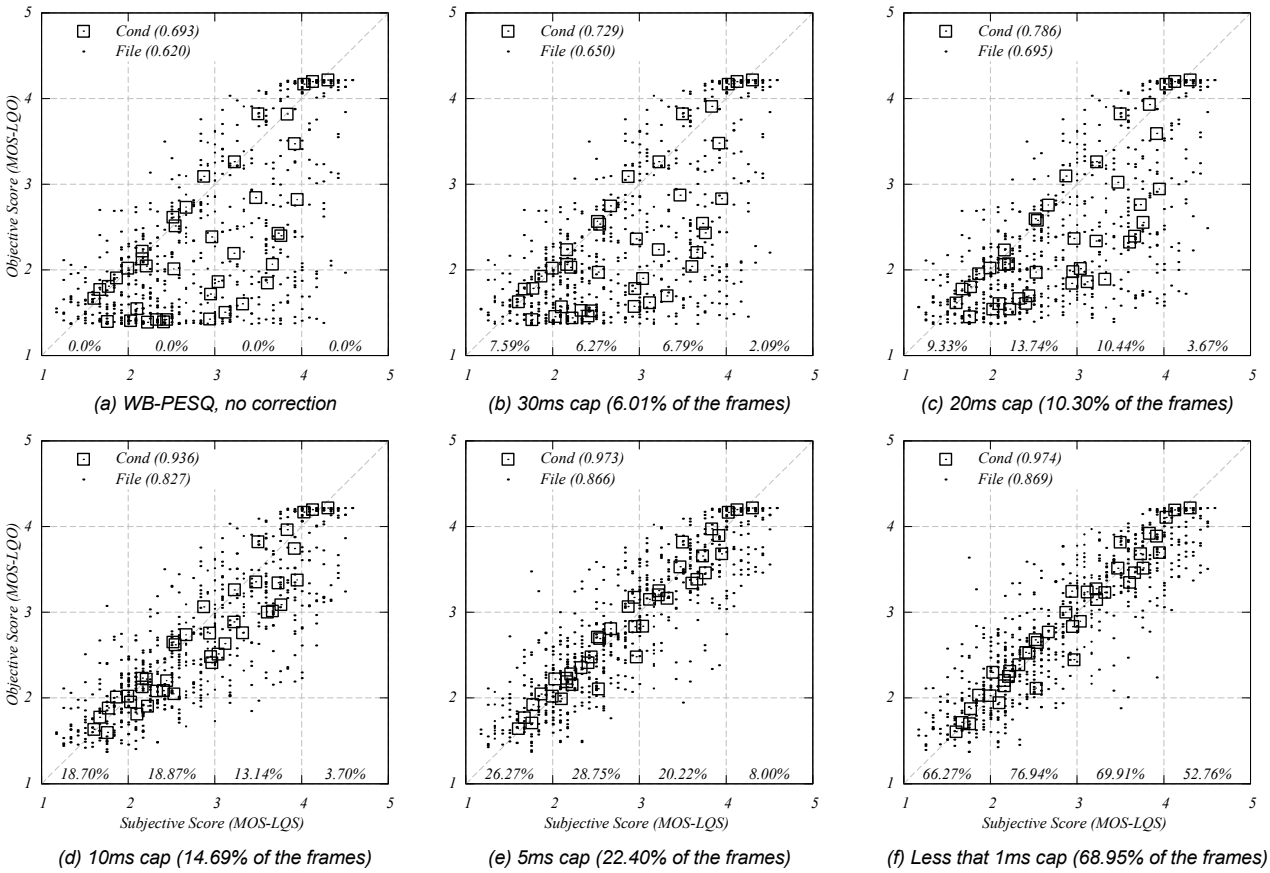

Fig. 3 WB-PESQ performance comparison against subjective score showing per file (dot) and per condition (square). Progressive capping of PESQ time-alignment error is shown in (a)-(f), with (a) representing the standard (poor) WB-PESQ result until (f) where the time-alignment error is effectively removed. Note also that the fraction of speech frames where the alignment was capped is shown in the x-axis legend e.g. in (b) 7.59% of speech frames for files exhibiting MOS-LQS between 1-2 were capped to a maximum of 30ms alignment error.

In order to predict such subjective tests, the time alignment in PESQ has to be modified and we found that the required time alignment accuracy to reach good performance is ±5ms. Further improvement may be gained by taking into account the impact of the time-warping on human perception of quality.

The need for an objective model capable of assessing the quality of newer VoIP systems is increasing with adoption of time-warping error correction techniques. This work was applied to correction techniques as found in software VoIP clients, but it should be noted that these techniques are also now being used in hardware clients and unified communication interfaces. A new activity for objective listening quality assessment was initiated at the ITU-T [11], partly for handling current PESQ limitations. Models capable of assessing time-warped signals, and therefore aligning signals accurately, are under active development; this paper defines the time-alignment accuracy that these models need to achieve.

7. REFERENCES

[1] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality", Aug. 2006

[2] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", Feb. 2001

[3] ITU-T Rec. P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO", Nov. 2003

[4] ITU-T Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs", Nov. 2005

[5] ITU-T Rec. P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications", May 2004

[6] L. Malfait, J. Berger, and M. Kastner, "P.563 – The ITU-T Standard for Single-Ended Speech Quality Assessment", IEEE Transaction on Audio, Speech and Language Processing, Vol. 14, No. 6, pp. 1924-1934, Nov. 2006

[7] ITU-T Rec. P.564, "Conformance testing for narrowband voice over IP transmission quality assessment models", Jul. 2006

[8] Global IP System, "Measuring Voice Quality", Dec. 2006, white paper available from http://www.gipscorp.com

[9] ITU-T Rec. P.830, "Subjective performance assessment of telephone-band and wideband digital codecs", Feb. 1996

[10] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment, part I—Time-delay compensation," J. Audio Eng. Soc., vol. 50, no. 10, pp. 755–764, Oct. 2002.

[11] ITU-T, "Requirement specification for P.OLQA", ITU-T Study Group 12, Working Party 2, Temporary Document 52, Jan. 2007