QUALITY ESTIMATION OF TIME-SCALE MODIFIED SIGNALS

Margus Muskat

Laboratory of Phonetics and Speech Technology
Institute of Cybernetics at Tallinn University of Technology
Tallinn, Estonia
margus@phon.ioc.ee

ABSTRACT

The paper describes some properties and problems related to Perceptual Evaluation of Speech Quality (PESQ) delay compensation mechanism and relations between time-scale modification and quality estimate. The evaluation of PESQ algorithm with two types of signal stretching (uniform resampling-based stretch and pitch-preserving stretch) was performed and the listening tests with human listeners were carried out. PESQ performance was also compared against the 3SQM algorithm. Experimental results indicate that performance of PESQ is not sufficient for high precision quality estimation of time-scale distortions.

1. INTRODUCTION

Principles of objective speech quality evaluation algorithms have initially been derived from knowledge about basic psychoacoustic relations and later been improved by complementing the algorithm by additional functionalities. Perceptual Evaluation of Speech Quality (PESQ) is the quality evaluation algorithm that can be viewed as improved version of Perceptual Speech Quality Measure (PSQM) algorithm where one of the improvements is delay compensation mechanism absent in PSQM. Subjective quality estimate is referred to as estimate of Mean Opinion Score (MOS) or sometimes as MOS Listening Quality Objective (MOS-LQO). Perceptual audio quality is traditionally assessed by MOS which is the arithmetic mean of subjective scores that are given by human subjects in listening tests. Intrusive evaluation and estimation of MOS can be described as a function

$$\mathbf{v},\mathbf{s}(\mathbf{v}) \rightarrow MOS, \qquad MOS \in [1,5] \tag{1}$$

where $\mathbf{v}$ is a reference voice signal vector and $\mathbf{s}(\mathbf{v})$ is a degraded output vector of evaluated signal processing system. Non-intrusive quality estimation, as described by the International Telecommunication Union recommendation P.563 (3SQM), is carried out by using only the degraded output vector:

$$\mathbf{s}(\mathbf{v}) \rightarrow MOS, \qquad MOS \in [1,5] \tag{2}$$

The fundamental idea behind PSQM and PESQ is comparing a reference signal and a degraded signal in the frequency domain. The comparison is carried out by dividing both signals into 32 ms frames with 50% overlap and compared frame by frame [2]. This method could lead to problems when the time-scale of the signal is distorted – unrelated frames could be compared and unfoundedly large differences could be measured. Therefore frame by frame comparison would be unsuitable when the time-scale of large amount of frames is altered. Time-scale distortions are common when coded voice packets are transported over network as different packets are subjected to different amounts of transmission delay. Jitter buffers are used to cope with small delay variations, but in the case of large variations some additional processing has to be carried out, typically some stretching is applied. Simplified description of the quality evaluation process is given by the following multi-stage transform

$$\mathbf{v},\mathbf{s}(\mathbf{v},\mathbf{d}) \rightarrow P_{REF},P_{DEG},\mathbf{d}_{EST} \rightarrow \mathbf{dist} \rightarrow MOS \tag{3}$$

where $\mathbf{v}$ is reference voice signal vector, degraded signal $\mathbf{s}(\mathbf{v},\mathbf{d})$ is stretched version of $\mathbf{v}$ that is processed according to delay vector $\mathbf{d}$, P_{REF} and P_{DEG} are matrixes estimating perceptual representation of the reference and degraded signal, $\mathbf{d}_{EST}$ is the estimate of the delay vector $\mathbf{d}$, and **dist** is the distortion estimate vector describing how large is the perceptual distortion of $\mathbf{s}(\mathbf{v},\mathbf{d})$ when compared to $\mathbf{v}$. There is no possibility to obtain true, i.e. subjective values of P_{REF}, P_{DEG} or **dist**. Column index of the perceptual representation matrix indicates position of the corresponding frame in the time scale, row index corresponds to the pitch scale in Bark and matrix element values represent loudness in the Sone loudness scale [1]. PSQM calculates distortion vector **dist** elements by column-wise comparing the perceptual representation matrixes P_{REF} and P_{DEG}, but PESQ uses estimated delay values to locate correct frame locations to be compared. As a result, PESQ is able to tolerate delay variations, but at least two problems can be pointed out. First, accuracy of the delay estimation is not perfect ($\mathbf{d}_{EST} \neq \mathbf{d}$) and that causes frame alignment errors. Second, if delay estimates were perfectly accurate, how would delay vector $\mathbf{d}$ be related to MOS? Current version of PESQ algorithm does not deal with this problem explicitly. The relation $\mathbf{d} \rightarrow MOS$ could be approximated by generating various delay vectors and analysing related opinion scores. To be more exact, this relation depends also to some extent on voice signal $\mathbf{v}$ and stretching method $\mathbf{s}$, which makes this situation even more complicated. As this problem has not been widely investigated, there are no traditional guidelines for approaching it. From linguistic point of view the variation of time-domain measures of voice signal could be interpreted as a result of some unknown accent, but there is no general linguistic model that would enable the estimation of MOS.

2. EXPERIMENTS AND MEASUREMENTS

Current section describes how quality estimate is related to stretch measures. One possible method of studying the relation $\mathbf{d} \rightarrow MOS$ is evaluating signals corresponding to all possible combinations of delay vectors $\mathbf{d}_i$. Signals can be stretched in various ways, in this case stretch is defined by three parameters – beginning moment of the region to be stretched b_i, length of the region l_j and the amount of stretch a_k, i.e. the ratio of stretched

length to the original length. Consequently, the delay vector **d** is determined by three measures b_i, l_j, a_k and the relation $\mathbf{d}(b_i, l_j, a_k) \rightarrow MOS$ could be studied by varying these measures. No linear relation between these variables is expected. There is no widely known experimental data set or simplified model that would indicate how stretch parameters are related to listeners' subjective opinions. Exhaustive study would produce large amount of signals to be evaluated by the listeners. For example, when 25 different beginning moments of the region, 20 lengths of the region, 20 different stretch amounts and 10 signals would be used, then the number of signals to be evaluated would be 25×20×20×10=100 000 and when the average duration of signals would be 10 seconds then total duration of signals would be approximately 278 hours. When using different stretching algorithms the number of possible combinations would be even higher. Therefore only a simplified study would be feasible – one measure would be varied while others would be fixed. Due to limited amount of evaluated signals only PESQ and 3SQM MOS estimates are currently available and presented in the following sections. Before accuracy of these estimates can be assessed and final conclusions can be drawn, MOS has to be obtained from listening tests. Relations between frames, utterances, stretch measures and delay compensation are presented in figure 1.

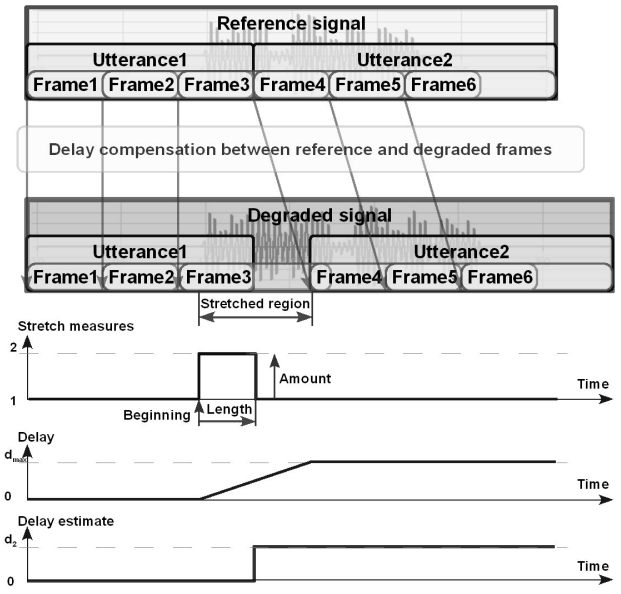

Figure 1: *Delay compensation during quality estimation.*

PESQ delay estimate $\mathbf{d}_{EST}$ can be described by the following multi-stage transform

$$\mathbf{v},\mathbf{s}(\mathbf{v},\mathbf{d}) \rightarrow \mathbf{env}_{REF},\mathbf{env}_{DEG},\mathbf{utt},\mathbf{d}_{env} \rightarrow \mathbf{d}_{EST} \quad (4)$$

where $\mathbf{env}_{REF}$ and $\mathbf{env}_{DEG}$ are energy envelopes of the reference and the degraded voice signals, **utt** describes division of the reference signal into utterances and $\mathbf{d}_{env}$ contains envelope-based delay estimates of the utterances with an approximate resolution of 4 ms. Initial utterance boundaries are determined by voice activity detector that measures energy of the reference signal and finally the number of utterances would be increased when delay changes during speech are detected by the utterance splitting procedure. Delay estimates are found from cross-correlation histograms of 64 ms long 75% overlapping signal frames located within the boundaries of the same utterance.

Figure 1 reveals a conceptual problem of any delay compensation algorithm – for every frame of the reference signal there is a corresponding frame of degraded signal, but due to stretch there could appear regions that are not related to any frame of the reference signal and therefore distortions of these regions would not influence the MOS estimate. As listeners perceive distortions of stretched regions that would be excluded by delay compensation algorithm, it would be appropriate to analyse perceptual properties of these regions as well. When locations of corresponding frames are searched by utilising detection of maximum cross-correlation, the delay compensation process would tend to exclude more distorted regions due to lower cross-correlation with the reference signal, and as a result, the delay compensation and MOS estimate would be somewhat arbitrary. Variation of signal duration is a natural property of speech, therefore relatively small change of duration should be perceived as a small decrease of quality. On the other hand, when content of the signal is musical and rhythmic then distortion of time-scale is more disturbing than in case of speech. As a result, in addition to delay vector the MOS depends on the content of the distorted signal. Due to aforementioned factors it would be difficult to develop a universal algorithm that could precisely estimate subjective importance of any time-scale distortion.

2.1. Variation of stretch amount

Figure 2 presents MOS estimates as a function of stretch amount for the case when pitch was not preserved due to resampling. As expected, maximum of the quality estimate is obtained when signal is not stretched. Otherwise, the estimate is mostly inversely proportional to the stretch amount that can be caused by misalignment of the reference and stretched signal frames. Figure 2 reveals that MOS estimate can be very sensitive to the stretch amount. Presented 3SQM estimates enable comparison of intrusive and non-intrusive quality estimation methods. Most noticeable difference is that 3SQM estimates do not form the peak around the least stretched region and 3SQM estimates are less sensitive to the stretch than PESQ estimates. Because of these contradictory results it can be concluded that both methods can not describe the relation correctly. Actual subjective opinions lie probably between these estimates, being less sensitive to stretch amount than PESQ and more sensitive than 3SQM.

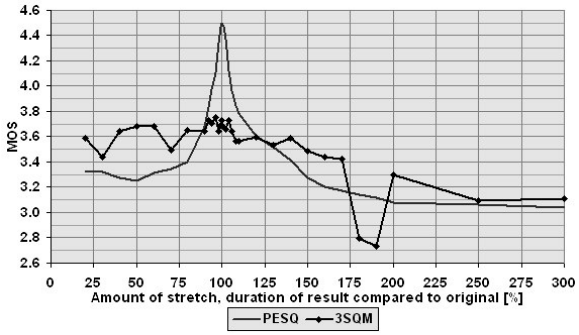

Figure 2: *MOS estimate as a function of resampling-based stretch amount.*

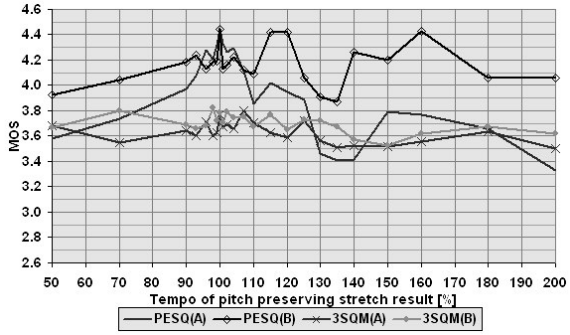

Figure 3: *MOS estimate as a function of pitch-preserving stretch tempo.*

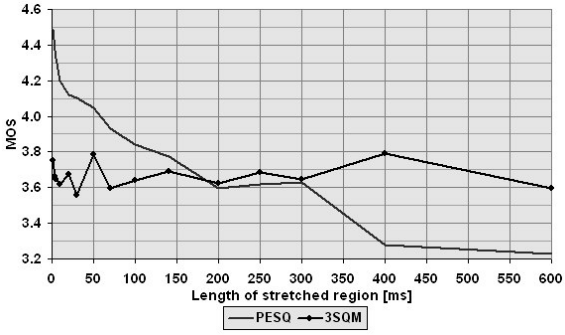

Figure 4: *MOS estimate as a function of stretch length.*

When pitch-preserving Pitch Synchronous Overlap Add (PSOLA) stretch is used, the relation between stretch amount and MOS estimate becomes more complex as figure 3 indicates. Figure 3 displays estimates obtained when the stretch was applied to two different regions (labelled as A and B) of the same signal. The most prominent difference is that PESQ estimates are not continuously diminishing but oscillating and 3SQM estimates reside in a relatively narrow range. Like in the case of stretch by resampling the actual MOS values could be expected to vary less than PESQ estimates and probably more than 3SQM estimates. These results indicate that different stretching algorithms and locations of stretch can cause different behaviour of quality estimation algorithm. Some high MOS estimates of signal B (around tempo 115% and 160%) seemed to contradict with informal subjective quality assessments, as there were some relatively easily detectable audible artefacts excluding the possibility of highest quality score. This can be caused by the fact that PESQ is not designed to cope with modulation effects. Results of PESQ and 3SQM are again contradictory – PESQ estimates are more sensitive to the stretch location than 3SQM estimates.

2.2. Variation of region length

Second measure that can be varied is the length of selected region. Figure 4 presents relation between the length of stretched region and corresponding MOS estimate when amount of uniform stretch is 90%, i.e. time-scale of the region is compressed. General trend of PESQ estimates is acceptable, except extreme sensitivity when the length of the stretched region is smaller than 20 ms. Informal subjective tests indicate that it is almost impossible to discern single stretches that are shorter than 20 ms.

Estimates of 3SQM are under the same conditions almost independent of the length of the region. Similar conclusion can be drawn as in the case of varying stretch amount – high sensitivity and lack of sensitivity are controversial properties that can not be valid simultaneously, therefore actual MOS values are probably somewhere between these results.

2.3. Variation of region location

The third measure that describes time-scale distortion is the beginning moment of stretch. Figure 5 presents relation between the beginning position of stretched region and the corresponding MOS estimate. MOS estimates were obtained by assessing two signals where male and female speakers pronounce the same sentence "She had your dark suit in greasy wash water all year". Due to slightly different speech tempo the word boundaries are not exactly aligned and only approximate timing of words is presented below the time scale. Length of the stretched region was 150 ms and the amount of stretch was two, i.e. the length was doubled. Distance between extreme estimates of 3SQM is about three times smaller than the distance between extreme estimates of PESQ. There is no obvious relation between PESQ and 3SQM estimates and most probably subjective opinion scores are somewhere in between these estimates. Just like in previous cases there is no widely accepted model that relates location of stretch to subjective opinion scores. When the reference signal is speech then the number of unique regions could be limited by the number of phonemes or classes of sounds that would be perceived similarly when time-scale of the region is modified.

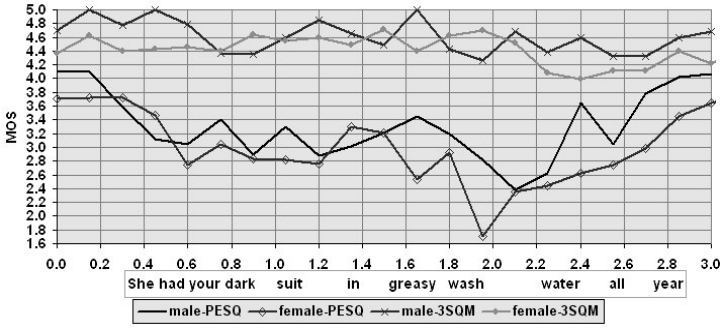

Figure 5: *MOS estimate as a function of stretched region's beginning.*

3. LISTENING TEST

For quality evaluation 77 signals were prepared, from these 32 signals were modified by time-scale distortion and 18 were distorted by stretches obtained by resampling. To avoid possibility that one large stretched region could be easily detectable and selection of the region could influence perception significantly, 25 regularly distributed 10 ms long regions were stretched. Informal listening tests indicate that single stretch of 10 ms long region is in most cases unperceivable when stretch amount is less than two. The reference signal was presented first, followed by the corresponding distorted signals that were presented in a single predetermined randomized order. Signals were evaluated by 14 listeners involved in telecommunication systems development. Two reference signals with male and female pronunciation of the same sentence were used. Figure 6 presents averaged MOS values, averaged MOS estimates (labelled as "MOSest") and differences between corresponding MOS values and MOS estimates. There is only one compressed signal per two or three expanded signals due to limited number of evaluated signals.

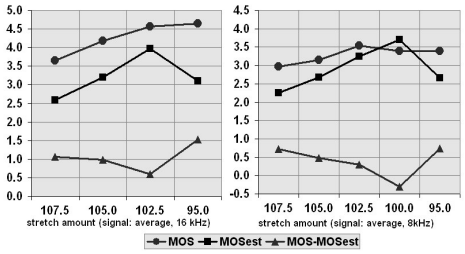

Figure 6: *Average of male and female speech MOS.*

These results point to an unexpected phenomenon – time stretch of a signal degrades subjective quality more than equivalent time-scale compression. In all occasions the MOS corresponding to the stretch amount of 105% is lower than MOS corresponding to the stretch of 95%. However, MOS estimates indicate that PESQ treats results of stretch and compression similarly. Largest difference between MOS and MOS estimates occurs when stretch amount is 95%. This phenomenon could be caused by time-domain post-masking and by the fact that frequency domain masking pattern is not symmetrical.

Measurements of Figure 6 also indicate that MOS variation caused by stretches decreased when sampling frequency was changed from 16 kHz to 8 kHz – difference between maximum and minimum average MOS in the 16 kHz case is 1.000 but 0.571 in the 8 kHz case. When a total subjective distortion would be divided into two distortion components – bandwidth distortion and stretch distortion, then MOS estimate can not be obtained by linear combination of different distortion components.

Figure 7 presents all 32 measurements of MOS and MOS estimates (MOSLQO) in the form of a scatter plot where the solid line is an approximation of the measurements by fourth order polynomial. Presented measurements indicate that PESQ tends to underestimate MOS when the time-scale of a signal is distorted – average of MOS is 3.455±0.461 and average of MOS estimates is 2.690, correlation coefficient is 0.685. Underestimation of high MOS values has been observed in various circumstances [3], as the time-scale distortion is not the only factor that can lead to underestimation of MOS.

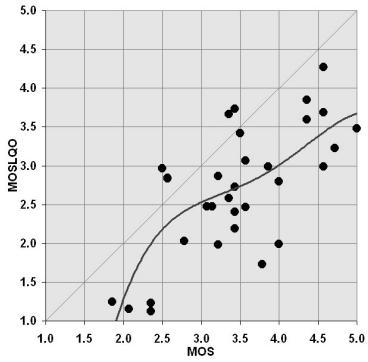

Figure 7: *Scatter plot of MOS and MOS-LQO.*

4. CONCLUSIONS

When compared to earlier objective perceptual quality measurement algorithms, PESQ is capable of handling quite large and complex time-scale distortions. However, experimental results indicate that performance of the time-delay compensation of PESQ is not sufficient for high precision MOS estimation of time-scale distortions. The most closely matching delay estimates are found when only relatively short regions are stretched and relatively long regions of signal are left unchanged. This is caused by the initial assumption that signals consist of utterances and delay changes occur between them.

There are at least two possible approaches how to improve MOS estimation accuracy in the case of time-scale distortions. The first approach is to increase accuracy of the delay estimates, i.e. the relation $\mathbf{d} \rightarrow \mathbf{d}_{EST}$; the other approach is to increase accuracy of the perceptual model $\mathbf{d} \rightarrow MOS$. Even when $\mathbf{d}_{EST} = \mathbf{d}$, current algorithm does not guarantee that MOS estimates would approach MOS. More listening tests should be conducted to obtain more opinion scores of time-scale distortions. During this study three parameters related to MOS in a manner that is not well known were identified: amount of stretch, length of stretched region, and position of stretch.

5. ACKNOWLEDGMENTS

This work was supported by Skype Technologies OÜ.

6. REFERENCES

[1] ITU-T, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Available at http://www.itu.int/rec/T-REC-P.862/, Accessed April 5, 2006.

[2] A.W. Rix, M.P. Hollier, A.P. Hekstra, J.G. Beerends, "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I – Time alignment," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 755-764, October 2002.

[3] S. Pennock, "Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) Algorithm," Available at http://wireless.feld.cvut.cz/mesaqin2002/, Accessed February 23, 2006.