CUSUM Application in Perceptual Speech Quality Control

Ahmad Zamani Jusoh, Roberto Togneri, Bijan Rohani, and Sven Nordholm, Member, IEEE

Abstract—One of the important aspects of the mobile communication industry is satisfying customers’ needs most economically. Indeed, customers expect good and consistent quality of service from the provider. In mobile telephony, this amounts to controlling the speech quality as “perceived” by the customers. Controlling the perceptual speech quality necessitates reliable measurement of the quality first, followed by exercising direct control of it. The end-users’ quality has traditionally been monitored and controlled based on radio link measurements such as the Signal Interference Ratio (SIR), Bit Error Rate (BER), or Frame Error Rate (FER). The ultimate measure of perceived speech quality is realized through subjective listening tests, but this is not practical for real-time day to day applications. In recent years, objective quality measurement algorithms have been developed to predict the subjective quality with considerable accuracy. The state of art in the International Telecommunication Union’s Telecommunication Standardization sector (ITU-T) recommendation for referenced objective quality measurement method is P.862 Perceptual Evaluation of Speech Quality (PESQ) model. These algorithms have yet to be applied for end user quality control in cellular networks using Statistical Process Control (SPC). In this paper, the application of Cumulative Sum (CUSUM) scheme, based on the PESQ algorithm for perceptual speech quality control is presented.

Index Terms—CUSUM, Perceptual Quality Measure, PESQ, SPC

I. INTRODUCTION

The speech quality was being monitored and controlled based on the conventional measurement such as Signal Interference Ratio (SIR), Bit Error Rate (BER) and Frame Error Rate (FER) measure is widely used is systems such as the 3rd Generation Universal Mobile Telecommunication System (3G UMTS) because it is considered as a good measure of speech quality. However, FER is not a perceptual measure of the speech quality. Furthermore, none of these non-perceptual measurements have been shown to estimate the speech quality with sufficient accuracy or reliability [1].

However, these parametric methods with their inferior performance are still commonly used. Since these methods lack accuracy in their prediction of the perceived speech quality, the service provider needs to cater for the worst case scenario in order to ensure that the quality expectations of almost all the customers are met; that is, the provider will have to unnecessarily expend more resources, such as transmission power and speech codec rate to prevent the speech quality from dropping below a certain acceptable limit. There are no constraints on the upper quality value. Therefore, often more than adequate quality is provided at the expense of the valuable resources. That is, the available methods do not control the perceptual quality directly but they do so indirectly through some relevant channel measures.

A truly perceptual quality measure is obtained when we analyze the received speech signal with a perceptual algorithm. There are two types of perceptual measurement methods: subjective and objective [2]. The subjective perceptual measure use a human subject or an end user in the communication systems named subjective perceptual speech quality measure. The subjective perceptual method is widely used but it is tedious, error-prone, expensive, and time consuming [2, 3]. By contrast, objective perceptual measure replaces the human subject by a computation model named Objective Speech Quality Measure in order to avoid the undesirable features of subjective tests.

Perceptual Evaluation of Speech Quality (PESQ) model is the state of art International Telecommunication Union’s Telecommunication Standardization sector (ITU-T) recommendation for referenced perceptual model measurement method. PESQ has been designed to improve on the previous objective methods. It is implemented commercially in testing devices and monitoring systems [4]. In this paper, the speech quality monitor and control based on PESQ algorithm will be discussed. The detail of PESQ is discussed at section II(C).

The control of perceptual speech quality using mechanisms such as power control and a “hybrid” control mechanism has been studied and applied before [5-7]. However, direct control approach using controlling tools such as Statistical Process Control (SPC) will be the first to be attempted. It is envisaged that the SPC which is a popular method in manufacturing and industrial process control, will be a promising method.

Statistical process control has been widely used in manufacturing and industrial quality control [8]. A statistical process control mechanism that has received much attention in the statistic literature and usage in industry in controlling the process mean is the Cumulative Sum (CUSUM) method. Furthermore, CUSUM scheme detects process shifts faster than any method [9]. Hence, in this paper a direct control
approach using CUSUM scheme to control the perceptual speech quality will be analyzed.

II. APPLICATION

Fig. 1 illustrates the concept of applying frame disturbance (FD) for controlling the perceptual quality of the received signal $y(n)$. The original signal and the degraded signal are required for the PESQ at the transmitter to calculate the FD for each frame. These calculated FDs can then be used for controlling some functions of the transmitter such as the transmission power, channel coding, or speech codec rate to maintain a required quality level. In the absence of $y(n)$ at the transmitting side, the PESQ must use an approximation of $y(n)$. One possibility for calculation of FD at the transmitting side, where the control is applied, is to use the Frame Quality Indicator (FQI) which was based on the Frame Erasure Pattern (FEP) information. This has been successfully applied before [7, 10, 11].

A. Original Speech

The original speech signals have been obtained from the International Telecommunication Union (ITU) database for voice quality measurement tests [12]. The signals were stored in files and prerecorded in 16-bit linear PCM (binary) format. Each of the constituent speech files contained prerecorded sentences of 8 seconds duration with approximately 50% speech and 50% silence intervals. 60 speech files are used in the simulation mixed male and female speakers and the silent parts were removed.

B. Adaptive Multi-Rate (AMR) Codec

The AMR speech codec is the standard codec for Universal Mobile Telecommunication Systems (UMTS). It was used in the analysis at the transmitter and receiver part. The AMR Codec is based on Algebraic Code Excited Linear Prediction (ACELP) technique [13, 14]. It encodes speech into frames of 20 ms duration and is rearranged into classes A, B, and C in decreasing order of their perceptual importance. There are eight codec modes and the number of bits in each frame varies depend on them. It is summarized in Table 1. The usage of AMR requires optimized link adaptation that selects the best codec mode to meet the local radio channel and capacity requirements. The codec mode is proportional with the quality of the speech, where the higher codec mode will result in better speech quality and vice versa [15].

<table>
<thead>
<tr>
<th>Codec Mode</th>
<th>Coded Rate (kb/s)</th>
<th>No. of bits per frame</th>
<th>No. of Class A bits</th>
<th>No. of Class B bits</th>
<th>No. of Class C bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4.75</td>
<td>95</td>
<td>42</td>
<td>53</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>5.15</td>
<td>103</td>
<td>49</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5.90</td>
<td>118</td>
<td>55</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>6.70</td>
<td>134</td>
<td>58</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>7.40</td>
<td>148</td>
<td>61</td>
<td>87</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7.95</td>
<td>159</td>
<td>75</td>
<td>84</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>10.2</td>
<td>204</td>
<td>65</td>
<td>99</td>
<td>40</td>
</tr>
<tr>
<td>7</td>
<td>12.2</td>
<td>244</td>
<td>81</td>
<td>103</td>
<td>60</td>
</tr>
</tbody>
</table>

TABLE 1

C. PESQ

Fig. 2 shows the block diagram of PESQ [4]. The algorithm requires both the original and the degraded speech signals to make a comparison. Signals are transformed frame-by-frame according to a perceptual model, which represents the human auditory system. The transformed signals are subtracted to calculate the FDs for each degraded frame. The FD represents the perceptual difference of the two signals. The FDs are aggregated over all frames and a mapping function is used to give a Mean Opinion Score (MOS) value for the degraded signal. The smallest period that PESQ can evaluate speech quality is 320 ms. This is too long for effective control of quality in the network. However FD is calculated every 16 ms. Even though 16 ms is too short for assessing the speech quality but it is suitable for control purposes. As such, FD is proposed for use as a perceptual metric to replace non-perceptual measures such as FER. The analysis of FD is discussed in section III (A).

The PESQ algorithm has been extensively used in measurement tools for accurate assessment of perceptual speech quality in modern telecommunication networks.

D. CUSUM as Controlling Tool

CUSUM is a fast and efficient statistical method for detecting parameter changes in a process. The process here is the observation of speech quality. In this study, tabular CUSUM will be applied for detecting the change of the mean of log(FD) of the speech signal. In current practice, formulas and equations presented by Montgomery were used [16].

With a log frame disturbance variable FD, which is assumed to be a normal distribution with mean $\mu$ and standard deviation $\sigma$, $\log(\text{FD}) \sim N(\mu, \sigma)$, the cumulative sums for...
detecting upward and downward shifts in the mean are calculated below:

\[ C_n^+ = \max[0, C_{n-1}^+ + \mu - (\mu_0 + K)] , \]
\[ C_n^- = \min[0, C_{n-1}^- + \mu - (\mu_0 + K)] , \]

Where \( \mu_0 \) is the target mean which is selected based on the
PESQ MOS target as shown in Table 2 at section III (A), \( \mu \) is the
process estimated mean, and \( K \) is usually called the reference value. It is chosen approximately halfway between
the target mean, \( \mu_0 \) and the out of control value of the process
mean \( \mu \), which one wishes to detect quickly.

If the shift is expressed in term of standard deviation as \( \mu_1 = \mu_0 + \delta \sigma \) (or \( \delta = |\mu_1 - \mu_0| / \sigma \)), then \( K \) is given by

\[ K = \frac{\delta}{2} \sigma = \frac{|\mu_1 - \mu_0|}{2} , \]

Using \( K = \frac{\delta}{2} \sigma \) is the appropriate choice for detecting a shift
approximately \( 1 \sigma \) in the process mean.

Notice that \( C^+ \) and \( C^- \) accumulative deviations from the
target value \( \mu_0 \) that are greater than \( K \), with both quantities, \( C^+ \) and \( C^- \) reset to zero upon becoming negative and positive
respectively. If either \( C^+ \) or \( C^- \) exceeds the decision threshold,
\( H \), the process is considered to be out of control.
The reasonable value for \( H \) is five times the process standard
deviation, \( \sigma \). \( H \) for the \( C^+ \) and \( C^- \) are known as CUSUM upper
limit and CUSUM lower limit respectively. The starting value
for the CUSUM, \( C_0^+ \) and \( C_0^- \) are taken to be zero.

The CUSUM variables \( C^+ \) and \( C^- \) are compared against
appropriate thresholds for detection of upward or downward
shifts. The thresholds are chosen based on the trade-off
between the responsiveness of the algorithm and the
probability of a false detection. Generally, thresholds which
lead to faster detection of a shift in the mean will also result in
a higher probability of false alarms or false detections.

Once the threshold is exceeded, a shift in the mean and
hence a change of speech quality outside of allowable range is
detected. Hence steps, such as change of AMR rate and power
control must be taken to restore the mean of its target range.
Once the action has been taken, the CUSUM must reinitialize
to zero.

E. Methodology

FD Analysis

By applying the FQI method, the synthesized speech signal
output \( y(n) \) with the PESQ MOS ranging 3.0 to 3.5 is
collected. In attaining more reliable FD distribution, 10 sets
with the same PESQ MOS are collected where each set
contains FD which was distributed by 10 speech files. Each
speech file in average contains 243 samples of FD. The
silence parts of the speech signal output \( y(n) \) were removed for
this analysis. The estimated mean and standard deviation of
the distribution of each PESQ MOS from 3.0 to 3.5 are
observed and recorded.

By applying the sample mean estimation theorem, the
estimated mean \( \log(FD) \), \( \mu_s \) for one set of speech file is given by

\[ E[\log(FD_n)] = \frac{1}{N} \sum_{n=1}^{N} \log(FD_n) = \mu_s , \]

Where \( N = 2430 \). Consequently, the estimated mean for 10
sets of speech files is given by

\[ E[\mu_s] = \frac{1}{N} \sum_{n=1}^{N} \mu_s = \mu_0 , \]

Where \( N = 10 \), and \( \mu_0 \) is the target mean of the \( \log(FD) \) that
will be used for CUSUM application.

CUSUM application

By applying the sample mean estimation theorem, the
estimated mean \( \log(FD) \) for one speech file is given by

\[ E[\log(FD_n)] = \frac{1}{N} \sum_{n=1}^{N} \log(FD_n) = \mu , \]

Where \( N = 243 \), and \( \mu \) is the mean of the \( \log(FD) \) that will be
used for CUSUM application.

The speech quality perceived among the end user is
different depending on their judgment of perception. This
CUSUM application is applied on an end user to end user
basis. Two cases will be applied with the CUSUM control
chart in this analysis. The quality of the speech was controlled
at PESQ MOS of 3.3. Hence CUSUM target mean, \( \mu_0 \) was
set to be 0.3559. A total of 50 sequence speech files are
simulated with the AMR initial speech codec rate set to 2. The
mean of \( \log(FD) \) for each speech file, \( \mu \) was applied to the
CUSUM control chart.

Case 1

The first 40 speech files are 3.3 PESQ MOS and the other
10 speech files are degraded speech files.

Case 2

The first 40 speech files are 3.3 PESQ MOS and the other
10 are the better grade of speech files.

III. RESULTS AND DISCUSSION

A. FD Analysis

The analysis of the FD shows that FDs have a log-normal
distribution for a given perceptual quality MOS. Table 2 shows that the mean of distribution of the \( \log(FD) \) are
increased with the degradation of the perceptual quality and
vice versa. The distribution suggests that for a given perceptual quality the FDs can have a wide range of values. Some large values can be tolerated while the overall quality remains the same.

<table>
<thead>
<tr>
<th>PESQ MOS</th>
<th>Target Mean, $\mu_0$</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>0.5671</td>
<td>0.0476</td>
</tr>
<tr>
<td>3.1</td>
<td>0.5340</td>
<td>0.0482</td>
</tr>
<tr>
<td>3.2</td>
<td>0.4692</td>
<td>0.0385</td>
</tr>
<tr>
<td>3.3</td>
<td>0.3559</td>
<td>0.0343</td>
</tr>
<tr>
<td>3.4</td>
<td>0.2557</td>
<td>0.0337</td>
</tr>
<tr>
<td>3.5</td>
<td>0.1550</td>
<td>0.0384</td>
</tr>
</tbody>
</table>

**B. CUSUM application**

The process standard deviation, $\sigma$ for the in controlled first 40 simulated speech files is 0.0878. $K$ was set to be $\frac{1}{2}\sigma$ and $H$ was set to be $5\sigma$. Therefore CUSUM upper and lower limit, $H$ was set to be 0.3512 and -0.3512 respectively. The process mean for the first 40 simulated speech files is 0.3575.

**Case 1**

Fig. 3. CUSUM control chart without controlling speech codec rate

Fig. 3 shows the CUSUM control chart without controlling the speech quality. The mean of process for the last 10 simulated speech files was increased to 0.4674. The out of control CUSUM was detected at the 43rd speech signal sample at the CUSUM upper limit which indicated there was a degradation of the speech signals.

**Case 2**

Fig. 5. CUSUM control chart without controlling speech codec rate

Fig. 5 shows the CUSUM control chart without controlling the speech quality. The mean of process for last 10 simulated speech files was decreased to 0.2446. The out of control CUSUM was detected at 44th speech signal sample but this time it occurred at the CUSUM lower limit. This indicated that this signal was beyond the quality which is needed by the end user or customer.
Their academic work.

GmbH for their permission in using PESQ software for use in

mode 1 starting at 45 rectified by decreasing the speech codec rate from mode 2 to

service level for all customers. This is achieved while maintaining satisfactory

effects to provide consistent perceived quality to

resources by providing just enough resources to meet required

provider and users. The provider can optimize the network

of this research will potentially benefit both the network

signals as required by the end users. Hence, it will help the

action at the transmitter to control the quality of the speech

such as FER can be replaced with FD of PESQ. By applying

result of the analysis shows that, the conventional parameter

unsatisfactory perceptual quality. To maintain a certain level

lead to inefficient utilization of resources and possibly

such speech codec rate should not be adapted on a frame-by-

result of FD analysis suggests that transmission parameters

is inversely proportional with the quality of the speech. The

Frame erasure pattern feedback

is used in Statistical Process Control," in

"Perceptual Quality Measurement and Control: Definition, Application and Performance," in 4th International


IV. CONCLUSION

Based on the FD analysis, it shows that the mean of log(FD) is inversely proportional with the quality of the speech. The result of FD analysis suggests that transmission parameters such speech codec rate should not be adapted on a frame-by-frame basis as is in the current practice. The current practices lead to inefficient utilization of resources and possibly unsatisfactory perceptual quality. To maintain a certain level of end-user perceptual quality, what is needed is to detect the shift in the distribution of FDs and take steps to rectify that such as controlled the transmission power, channel coding or speech codec rate that was being applied in this analysis. The result of the analysis shows that, the conventional parameter such as FER can be replaced with FD of PESQ. By applying this new parameter to the CUSUM scheme will allow faster action at the transmitter to control the quality of the speech signals as required by the end users. Hence, it will help the provider in optimizing the network resources. The outcomes of this research will potentially benefit both the network provider and users. The provider can optimize the network resources by providing just enough resources to meet required levels of service beside provide consistent perceived quality to the customers. This is achieved while maintaining satisfactory service level for all customers.

ACKNOWLEDGMENT

The authors wish to thank, and acknowledge OPTICOM GmbH for their permission in using PESQ software for use in their academic work.

Fig. 6. Apply CUSUM in controlling speech codec rate

Fig. 6 shows the excess quality of the speech signals was rectified by decreasing the speech codec rate from mode 2 to mode 1 starting at 45th speech signal sample.