Adaptive Control of Perceptual Speech Quality in Modern Wireless Networks

Bijan Rohani, Behrooz Rohani*, Manora Caldera*, Hans-Jürgen Zepernick***
{behrooz, bijan, caldera}@watri.org.au; hans-jurgen.zepernick@bth.se
**Western Australian Telecommunications Research Institute, Perth, AUSTRALIA*
*** Blekinge Institute of Technology, Ronneby, SWEDEN*

Abstract

The speech quality in modern wireless networks has been traditionally measured by metrics which are derived from radio link measurements. The indirect measurement of speech quality based on such metrics is unreliable and often precious network resources are sacrificed to make up for this. In this paper, a perceptual metric is considered for direct measurement of speech quality and the framework for adaptive control of the perceptual speech quality is presented.

Keywords

Perceptual quality, Speech quality control

1 Introduction

The "perceived" speech quality is considered as an important component of the overall quality of service (QoS) at the receiving end of a communication link. Speech quality has been traditionally measured and controlled based on radio link parameters such as the signal-to-noise ratio (SNR), bit error rate (BER), and frame erasure rate (FER). These can be classified as parametric or indirect speech quality evaluation (SQE) methods. In general, such methods give unreliable estimates of speech quality and as such lead to inconsistencies in the speech quality and ultimately cause dissatisfactions with the QoS. In this case, precious radio resources are often sacrificed in order to improve speech quality [1].

The true measure of speech quality is obtainable through subjective listening tests [2]. However, these are not practical for most applications. As such, objective SQE algorithms have been developed and improved over the years. These algorithms are based on the psychoacoustic model of the human hearing system, and closely predict the subjective quality of speech signals. A prominent example of such algorithms is the perceptual evaluation of speech quality (PESQ) which was recommended by the International Telecommunications Union (ITU) as the standard for objective speech quality estimation [3]. This algorithm measures the speech quality by comparing the speech signal, possibly distorted due to codec and network impairments, with the original speech signal after necessary time alignments and signal level adjustments. Even though PESQ has provided the network operators with a tool to accurately monitor the speech quality in their networks, its application has been restricted to off-line processing. This has been due to the PESQ requirement to have both the original and the distorted signals available at the point of quality measurement. This however is a requirement that cannot be met in real-time applications. As such, the ITU has standardized a single ended assessment model (SEAM) [4] which provides a prediction of the subjective speech quality based only on the distorted signal. This model is however substantially more complex and less reliable compared with PESQ.

An alternative approach for SQE is discussed in this paper. This approach is a mix between parametric and psychoacoustic methods with comparable performance and complexity to the PESQ. Subsequently, a frame work for adaptive control of perceptual speech quality in modern wireless networks is presented.

This paper is organized as follows. The frame erasure pattern feedback method is described in Section 2. This method incorporates the PESQ algorithm to reliably measure the perceptual speech quality in a wireless network in real-time. This is followed by adaptive perceptual speech quality control in Section 3. Here, a psychoacoustic metric, namely the frame disturbance, is derived from the PESQ, and its characteristics are discussed. Then a statistical process control approach is proposed for the control of the speech quality based on these characteristics. The paper is ended with conclusions in Section 4.

2 Frame Erasure Pattern Feedback Method

Modern wireless systems widely employ speech codecs whose various bits at the output have unequal perceptual importance. Therefore, these bits are often rearranged according to their perceptual importance so that unequal error protection can be applied against transmission errors. In this way, those bits that are more important for the reconstruction of the speech are protected more effectively. For example, in the 3G Universal Mobile Telecommunication System (UMTS) adaptive multi-rate (AMR) codec, the encoded speech bits within a speech frame are rearranged into Class A, Class B, and Class C bits in decreasing order of their perceptual importance [5]. In a typical implementation, Class A bits are protected by rate-1/3 convolutional coding (CC), Class B bits with rate-1/2 CC, whereas Class C bits may be left unprotected. Class A bit errors can cause undesirable artifacts in the reproduced speech, whereas Class B and Class C bit errors cause less severe degradation of speech quality. For this reason, in addition to applying extra error protection to Class A bits, an error-concealment mechanism is provided at the speech decoder to mask the undesirable effects of erroneous Class A bits [6]. Class A bit errors result in erasure of the corresponding frame. The error-concealment mechanism subsequently replaces the erased frame with one that it calculates from the previous frames with correct Class A bits. Frame erasure (FE) is decided based on the so-called bad frame indicator (BFI). This is a binary flag associated with every received speech frame by the channel decoder to indicate the quality of Class A bits. A frame is said to be "bad" if the Class A bits contain errors after the channel decoding, otherwise the frame is considered to be "good".

It should be noted that the perceptual quality in the presence of bad frames is affected by the similarity of the contents of the substitution frame, calculated by the error concealment mechanism, with the original contents of the erased frame. This has been demonstrated in Fig. 1 where a single frame erasure has been assumed during an 8-second (400 frames) speech file. The position of the single frame erasure has been varied and the resulting perceptual quality has been calculated based on the PESQ algorithm. As it can be seen from Fig. 1, a range of speech qualities, measured in mean opinion score (MOS), was observed. Although majority of frame erasures resulted in negligible degradation in quality, some cases with a drop in quality by as much as 0.5 of a MOS unit were also observed.

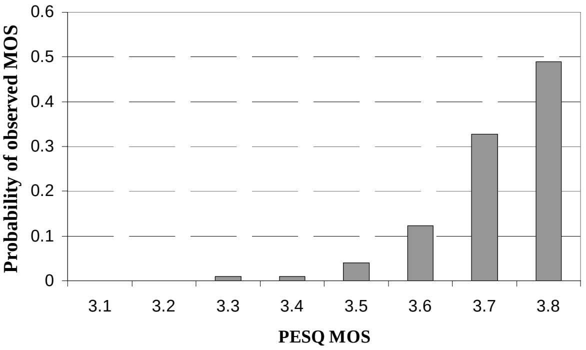

Fig. 1 The probability of experiencing a given speech quality with a single frame erasure. Note: PESQ MOS measured under error-free condition was 3.8.

In practice, multiple frame erasures may affect a short segment of the received speech signal. In such an event, the temporal distribution of frame erasures can influence the effectiveness of error concealment and the resulting speech quality. Heuristically, consecutive erasures can affect the speech quality more adversely than if the erasures are distant from each other. The frame erasure pattern (FEP) feedback scheme [7] has been based on these properties of the error-concealment mechanism to facilitate real-time measurement of speech quality of the receiving side at the transmitter. The block diagram of this method is shown in Fig. 2. In this case, the 20 ms speech frames $\mathbf{x}^{(n)}$; n=1, 2, …, are encoded by the AMR speech encoder. The encoded frames $\mathbf{x}_Q^{(n)}$ are then processed in a chain of physical layer stages including channel coding, interleaving, and modulation. The entire physical layer processing in the transmitter and receiver has been represented by Tx and Rx functions in Fig. 2. It should be noted that in general the physical layer can be that of any modern wireless network.

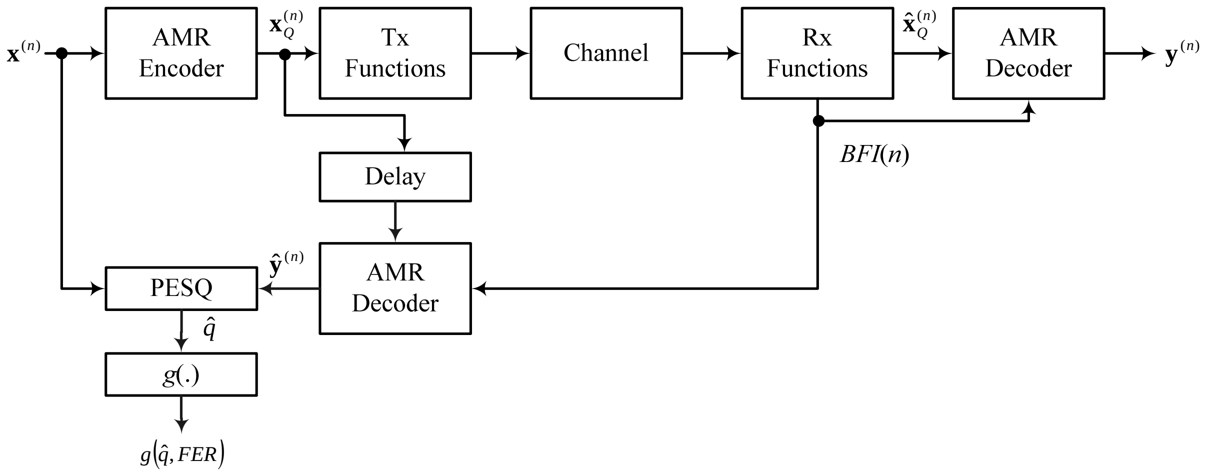

Fig. 2 The simplified block diagram of the FEP feedback method.

Subsequently, the corresponding recovered frames $\hat{\mathbf{x}}_Q^{(n)}$ at the receiver are speech decoded to render the output speech frames $\mathbf{y}^{(n)}$. The AMR decoder relies on the bad frame indication flag *BFI*(*n*) for individual frames $\hat{\mathbf{x}}_Q^{(n)}$ in order to invoke the error concealment procedure if necessary. The sequence of 1-bit flags *BFI*(*n*), *n*=1, 2, … represents the FEP. This is fed back to the transmitter, where in conjunction with the original transmitted frames $\mathbf{x}^{(n)}$, an approximation $\hat{\mathbf{y}}^{(n)}$ of the receiver speech signal $\mathbf{y}^{(n)}$ is constructed according to

$$\hat{\mathbf{y}}^{(n)} = Q^{-1}\left[\mathbf{x}_Q^{(n)}, BFI(n)\right] \tag{1}$$

Here, Q^{-1} represents the AMR speech decoding functions including the error concealment procedure. Note that, in general, Q^{-1} is also a function of the AMR speech encoding rate not shown here.

The perceptual quality $\hat{q}$ of the approximated output for a sequence of N frames $\left\{\hat{\mathbf{y}}^{(n)}\right\}_{n=1}^{N}$ is calculated based on the perceptual distance D of this signal with the corresponding sequence of the original signal frames using the PESQ algorithm, according to

$$\hat{q} = D\left[\left\{\mathbf{x}^{(n)}\right\}_{n=1}^{N}, \left\{\hat{\mathbf{y}}^{(n)}\right\}_{n=1}^{N}\right] \tag{2}$$

Since $\hat{\mathbf{y}}^{(n)}$ does not account for the so-called residual BER at the receiver, a mapping function g has been used to correct for this omission. In this case, the perceptual quality q corresponding to the receiver output $\mathbf{y}^{(n)}$ is given as

$$q = g(\hat{q}, FER) + \varepsilon \tag{3}$$

where *FER* represents the average frame erasure rate during the measurement period, and ε is the estimation error due to FEP feedback scheme. Once again, it is noted that g is also a function of the AMR coded rate but for convenience it has not been shown in (3).

The performance of the FEP feedback scheme has been investigated for the 3G UMTS through computer simulations. The results have shown close agreement between the estimated quality $g(\hat{q}, FER)$ and the actual quality q under various channel conditions and AMR codec rates. The correlation coefficient between q and $g(\hat{q}, FER)$ has been reported to fall in the range of 0.82 and 0.93, while the corresponding root-mean-square estimation error ranged from 0.08 to 0.15 of a unit on the MOS scale [7].

3 Adaptive Perceptual Speech Quality Control

3.1 Frame Disturbance

As discussed in Section 1, parametric SQE methods only provide indirect measures of quality and as such are not suitable for reliable speech quality control. In order to exercise direct control over the delivered speech quality in wireless networks, it is necessary to apply psychoacoustic measures.

The FEP feedback scheme shown in Fig. 2 may seem appropriate for application with adaptive control of perceptual speech quality. However, it is noted that the PESQ requires input speech segments of at least 160 ms in duration for reliable processing. In practice, this will give rise to an unacceptably long delay to react to channel changes while trying to maintain a desired perceptual quality at the output.

The PESQ MOS is derived from the so-called frame disturbances (FD). These are effectively the perceptual distances between the distorted signal and the original signal calculated on a frame-by-frame basis [3]. Calculation of FD is complex and mimics the sound wave transformations in human inner ear. It involves calculation of the short-term signal power spectral density (PSD) followed by frequency and amplitude transformations characterizing human ear's psychoacoustic model. It is therefore plausible to adopt FD as a psychoacoustic replacement for parametric SQE measures, such FER, for the control of the perceptual speech quality. In this case, the configuration shown in Fig. 2 can be used for calculation of FD and ultimately control of the output speech quality.

The FD for a sufficiently long period of speech signal transmitted under a given channel condition is randomly distributed. It has been found through simulations that the distribution of FD can be modeled with a log-normal distribution. An example of a typical FD distribution is shown in Fig. 3. In this case, the measured MOS was 3.5. It has been found that, in general, the mean of the distribution increases with degradation of the perceptual speech quality.

The distribution suggests that for a given perceptual quality the FD can have a wide range of values. Some large values can be tolerated while the overall quality remains the same. This implies that transmission parameters such as power should not be adapted on a frame-by-frame basis as is 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, all that is needed is to detect a shift in the distribution of FD and take steps to rectify that. One way that this can be achieved is through statistical control approach.

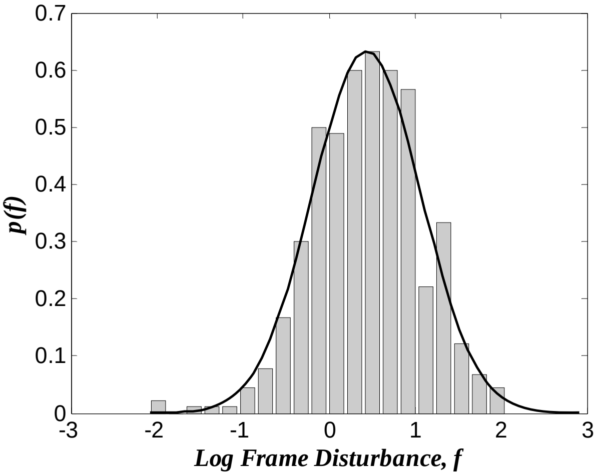

Fig. 3 The probability density function p(f) of frame disturbances for a typical speech signal approximately follows a log-normal distribution.

3.2 Statistical Control

Statistical process control (SPC) has been in use for quality control in manufacturing and chemical engineering [8]. In SPC, the quality of a process is monitored over time to ensure that it stays in a state of statistical control. Such a state of control exists if key process variables remain close to their desired values. The only source of variations in this case is "common-cause" variation that affects the process all the time and is unavoidable. Once in the state of control, no action is taken in response to common-cause variations. However, in the event that the state of control is disturbed due to occurrence of "special-cause" events, appropriate action must be taken to restore state of control.

Because of the random nature of FD, statistical process control can be adopted for speech quality control. In this case, the perceptual speech quality is the process to be kept in the state of statistical control. The process variable is the measured FD. When the perceptual speech quality is in a state of control, the measured FD follows a log-normal distribution around a mean value that determines the desired perceptual quality of the speech. The variations of the measured FD under such circumstances are considered common cause and warrant no action. However, if the mean of the process drifts from its desired value, i.e. a special cause is detected, the necessary action is triggered in order to restore the process to its state of control.

A popular SPC method for detection of small drifts in the process mean is the cumulative sum (cusum) [8] scheme. Let f_n; n=1, 2, … be a sequence of i.i.d. random variables representing measured FDs. The low-side and high-side cusum schemes L_n and H_n, respectively, with initial values $L_0 = H_0 = 0$ are defined as

$$L_0 = 0, \quad L_n = \min\left[0, L_{n-1} + \left(f_n - \mu\right)\right] \qquad n = 1,2,... \qquad (4a)$$

and

$$H_0 = 0, \quad H_n = \max\left[0, H_{n-1} + \left(f_n - \mu\right)\right] \qquad n = 1,2,... \quad (4b)$$

where μ is the desired process mean.

A downward shift in the process mean is detected as soon as the low-side cusum falls below a threshold, i.e. $L_n \le -h$ where h is a positive constant. Similarly, an upward shift in the mean is signaled by the high-side cusum when $H_n \ge h$. The cusum is reinitialized once the threshold is crossed.

An example of the cusum for detecting an upward shift in the process mean is shown in Fig. 4. The mean of the process was 0.5 for the first 100 FD measurements and it was changed to 1.0 for the next 100 measurements. A threshold value of h=4.6 was used for this example. With this threshold, a change in mean is detected on average within 12 samples. In this case, the shift was detected after 14 samples after the shift had occurred.

4 Conclusion

In this paper, it was explained that parametric SQE methods, which have been traditionally used in wireless network for speech quality estimation, are unreliable. Application of such methods for speech quality control leads to inefficient radio resource management and possibly end-user dissatisfaction with the QoS. Methods based on the psychoacoustic model of human hearing are more reliable but not suitable for real-time applications. The state of the art SEAM which is useful for real-time monitoring of network speech quality is not sufficiently reliable for speech quality control. In this case, the FEP feedback method that has comparable reliability to the PESQ method was discussed for potential application in speech quality control. The FD, which is derived from PESQ algorithm, was proposed as a perceptual replacement for the conventional quality metrics such as the FER. This could then be used in conjunction with SPC cusum scheme for adaptive control of the perceptual speech quality in modern wireless networks.

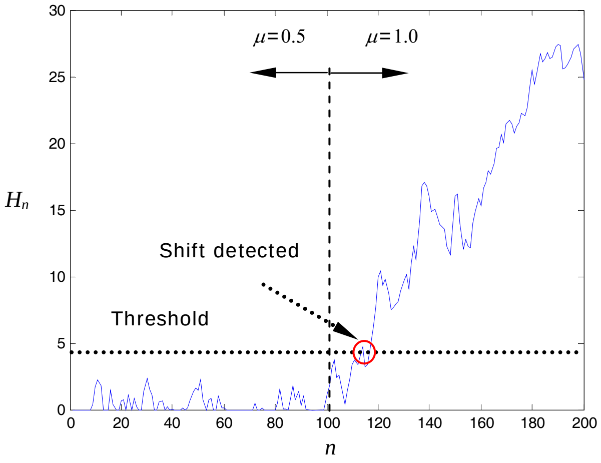

Fig. 4 An example of the high-side cusum detecting an upward shift in the process mean μ.

5 Acknowledgement

This work is partly funded by the National ICT Australia. National ICT Australia is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

6 References

[1] B. Rohani, B. Rohani, M. Caldera, and H.-J. Zepernick, "Benefits of Perceptual Speech Quality Metrics in Modern Cellular Systems", Proc. IEE Electronics Letters, vol. 42, no. 21, Oct. 2006

[2] "ITU-T Recommendation P.800: Methods for Subjective Determination of Transmission Quality," Aug. 1996.

[3] "ITU-T Recommendation 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.

[4] "ITU-T Recommendation P.563: Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications", May 2004

[5] "3G Technical Specification TS 26.101: AMR Speech Codec Frame Structure", Mar. 2002.

[6] "3G Technical Specification TS 26.091: AMR Speech Codec; Error Concealment of Lost Frames", Mar. 2001.

[7] B. Rohani, B. Rohani, and H. J-. Zepernick, "Feedback Method for Real-time Perceptual Quality Estimation", Proc. IEE Electronics Letter, vol. 40, no. 14, Jul. 2004.

[8] G. Box and A. Luceño, "Statistical Control by Monitoring and Feedback Adjustment", Wiley-Interscience, 1997.