THE INVESTIGATION OF FRAME DISTURBANCE (FD) IN PERCEPTUAL EVALUATION SPEECH QUALITY (PESQ) AS A PERCEPTUAL METRIC

Ahmad Zamani Jusoh[1], Roberto Togneri[2], Sven Nordholm[3], Nadzril Sulaiman[4] and
Muhamad Haziq Khairolanuar[1]
[1]Faculty of Electrical and Computer Engineering, International Islamic University of Malaysia (IIUM), Kuala Lumpur, Malaysia
[2]School of Electrical, Electronic and Computer Information, University of Western Australia, Crawley, Australia
[3]School of Electrical Engineering and Computing, Curtin University of Technology, Perth, Australia
[4]Faculty of Mechatronics Engineering, International Islamic University of Malaysia (IIUM), Kuala Lumpur, Malaysia
E-Mail: azamani@iium.edu.my

ABSTRACT

Satisfying customers' needs economically is one of the important aspects in mobile communication industry. Provider should cater a good and consistent quality of service as expected by the customers. Hence, it is amounts to controlling the speech quality perceived by the customers. However, to control the speech quality, the reliable measurement of the speech quality must be determined first, then exercising direct control of it. Traditionally, the quality of speech at the end user has 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 truly perceive speech quality is measure through subjective listening tests. However, this method is not practical for real-time day to day applications. Newly, objective quality measurement algorithms have been developed to replace the subjective quality measurement with considerable accuracy. P.862 Perceptual Evaluation of Speech Quality (PESQ) model is the state of art on the referenced objective measurement method in the International Telecommunication Union's Telecommunication Standardization Sector (ITU-T). PESQ is calculated based on the so-called Frame Disturbance (FD). In this paper FD is investigated as a perceptual metric for control of speech quality in modern communication systems replacing the conventional metrics.

Keywords: perceptual, quality measure, PESQ, FD.

INTRODUCTION

To date, the conventional measurement methods had applied in communication systems to monitor and control the speech quality such as Signal Interference Ratio (SIR), Bit Error Rate (BER), and Frame Error Rate (FER). Among them, FER measure is widely used in systems for instance the 3G UMTS (Universal Mobile Telecommunication System) because it is considered as a good measure of speech quality. However, FER is not a perceptual measure of speech quality. In addition, none of these non-perceptual measurements have been demonstrated to estimate the speech quality with acceptable accuracy or reliability [1-3].

Despite with inferior performance, these parametric methods are still commonly used in communication systems. Since the lack accuracy in their estimation of the perceived speech quality, the service provider needs to cater for the worst case scenario to almost all the customers to ensure the speech quality are meet the expectations. Therefore, the provider will have to unnecessarily dispense more resources, such as transmission power and speech codec rate to avoid the speech quality from decreasing below a certain threshold. There are no compulsions on the upper quality value. Therefore, often more than sufficient quality is afforded at the expense of the valuable resources. The scenario happened because the available methods do not control the perceptual quality directly but they just control indirectly through some relevant channel measures.

A truly perceptual quality measure can be obtained by analyzing the received speech signal with a perceptual algorithm. There are two types of perceptual measurement methods: subjective and objective [3]. The subjective perceptual measure uses 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 [3-5]. On other hand, objective perceptual measure replaces the human subject by a computation model named Objective Speech Quality Measure in order to prevent the undesirable features of subjective tests.

The state of art International Telecommunication Union's Telecommunication Standardization sector (ITU-T) recommendation for referenced perceptual model measurement method is Perceptual Evaluation of Speech Quality (PESQ) model. PESQ is the improved model of the previous objective methods. It is implemented commercially in monitoring systems and testing devices [6].

However, the smallest period that PESQ can evaluate the speech quality is 320 ms [7]. Even though this or longer periods may be suitable for monitoring speech quality, it may be too long for persuasive control of quality in the communication systems. As such, it will be necessary to investigate metrics which can be calculated faster than 320 ms for application in controlling the quality.

The PESQ is calculated based on so-called "Frame Disturbance", (FD) which is effectively the perceptual distance between the reference and the distorted speech signals [7]. The FD is calculated every 16 ms. Although 16 ms is too short for assessing speech quality it is suitable for control purposes. Hence, tt is proposed that the FD is investigated as a perceptual metric for control of speech quality in modern networks replacing conventional metrics such as SIR, BER and FER which will give more accuracy in predicting the perceived speech quality. That will be the key contribution of the research. The preliminary result of FD analysis was conducted in [8]. In this paper, the analysis of FD in details will be presented.

PESQ

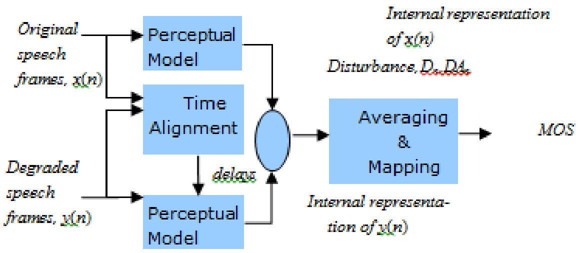

Original speech frames, x(n)
Perceptual Model
Time Alignment
Degraded speech frames, y(n)
delays
Perceptual Model
Internal representation of x(n)
Disturbance, D_n,DA_n
Averaging & Mapping
MOS
Internal representa-tion of y(n)

Figure-1. PESQ block diagram.

The block diagram of PESQ is shows in Figure-1 [6]. PESQ requires both the original and the degraded speech signals to make a comparison. Signals are transformed frame-by-frame based on the perceptual model, which represents the human auditory system. The transformed signals are subtracted to calculate the FD for each degraded frame. The FD represents the perceptual difference of the two signals and is aggregated over all frames. A mapping function is used to give a Mean Opinion Score (MOS) value for the degraded signal. The smallest period that PESQ can assess speech quality is 320ms. This is too long for persuasive control of quality in the network. However FD is calculated every 16ms. Although 16ms 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 take over non-perceptual measures such as FER.

FD

The sign difference between the distorted and original loudness density in PESQ is called the raw disturbance density. The minimal of the original and degraded loudness density is computed for each time frequency cell. This results in a disturbance density as a function of time (window number *n*) and frequency, $D(f)_n$. As PESQ involves the asymmetry effect processing, the asymmetrical disturbance density, $DA(f)_n$ is also aggregated [6, 7].

The disturbance, D_n and the asymmetrical disturbance, DA_n are calculated by a non-linear average as below:

$$D_n = M_n \sqrt[3]{\sum_{f=1,\ldots Nb} (|D(f)_n|W_f)^3} \tag{1}$$

$$DA_n = M_n \sum_{f=1,\ldots Nb} (|DA(f)_n|W_f)^3 \tag{2}$$

M_n is a multiplication factor which is equal to *1/(power of original frame + 10⁵/10⁷)*$^{-0.04}$ and Nb is the number of bark bands. W_f is a series of constants which are proportional to the width of the modified Barks bins. This results in disturbance and asymmetrical disturbance signals that represent how distorted the speech is during a very short period of time (16ms). The linear combination of disturbance and asymmetrical disturbance values will result in the final disturbances which are referred to as FD_n throughout this paper as

$$FD_n = D_n + DA_n \tag{3}$$

FD Analysis

Figure-2 illustrates the simulation model for FD analysis. In order to calculate the FD for each frame, the original signal and the degraded signal are required for the PESQ at the transmitter. Due to the absence of *y(n)* at the transmitting side, the PESQ must use an approximation of *y(n)*. Frame Quality Indicator (FQI) which was based on the Frame Erasure Pattern (FEP) information is applied for that purpose. This has been successfully applied before [9-11].

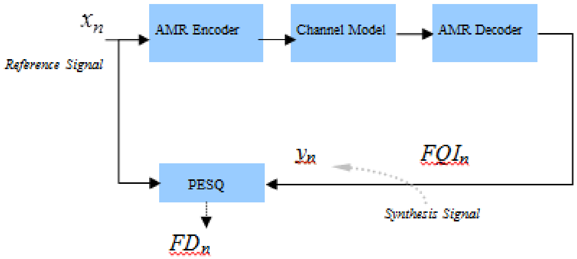


Figure-2. Simulation model for FD analysis.

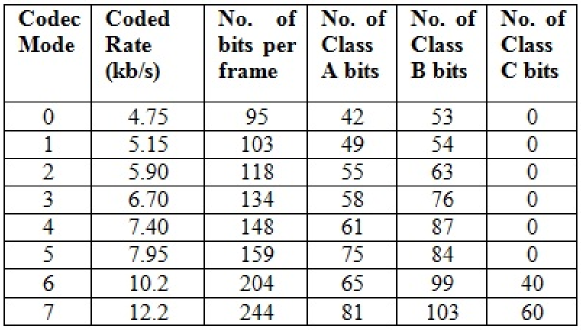
Input Speech File and Speech Codec

Input speech samples used in the analysis are from the ITU database for speech quality measurement test [12].The signals were prerecorded in 16-bit linear PCM (binary) format and stored in files. Each of the constituent 8 seconds duration speech files contained prerecorded sentences of with approximately 50% speech and 50% silence intervals. However, FD is calculated with the silence periods removed to ensure only the active speech is considered in this application. The AMR speech codec is the standard codec for UMTS It was used in the analysis at the transmitter and receiver part [13].

Adaptive Multi-Rate (AMR) Codec

The AMR speech codec is the standard codec for Universal Mobile Telecommunication Systems (UMTS). The AMR Codec is based on Algebraic Code Excited Linear Prediction (ACELP) technique [14, 15]. It was used in the analysis at the transmitter and receiver part. It encodes speech into frames of 20 ms duration. It is reorganize 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 which 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-1. Number of bits in Classes A,B, and C for each AMR codec mode [13].



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

METHODOLOGY

The degraded speech signal with PESQ MOS ranging 3.0 to 3.5 is collected and saved. In achieving more reliable FD distribution, for each PESQ MOS, 10 sets are collected where each set contains FD which presented by 10 speech files. Each speech file on average contains 243 samples of FD calculations. The silence parts of the speech signal output were removed for this analysis. The simulation model is as Figure 2. The estimated mean and standard deviation of the distribution of each PESQ MOS from 3.0 to 3.5 are examined and recorded. Then, the relative frequency of log (FD_n) are plotted.

By applying the sample mean estimation theorem, the estimated mean $\log\left(\mathrm{FD}_n\right), \mu_S$ for one set of 10 speech file is given by

$$E[\overline{log\,(FD_n)}] = \frac{1}{N}\sum_{n=1}^{N} log\,(FD_n) = \mu_s\,, \tag{4}$$

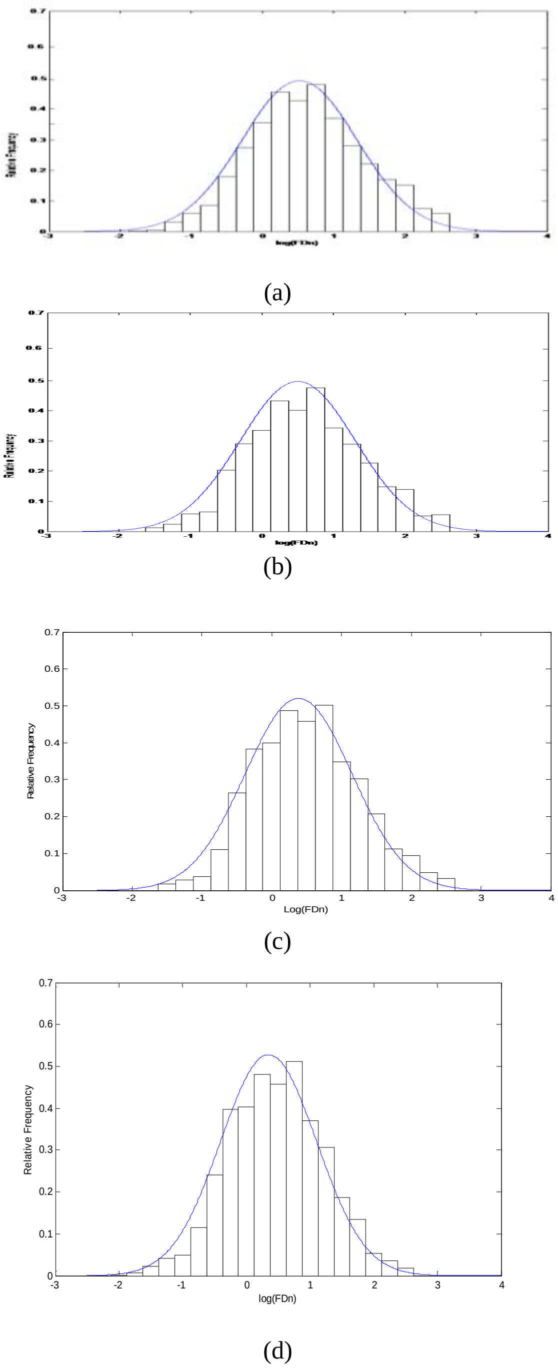
where N = 2430 for 10 speech files. Consequently, the estimated mean for all 10 sets of speech files is given by

$$E\,[\,\overline{\mu_s}\,] = \frac{1}{N}\sum_{m=1}^{M} \mu_{sn} = \mu_0\,, \tag{5}$$

where M = 10, and μ_0 is the target mean of the $\log\left(FD_n\right)$.

RESULTS AND DISCUSSIONS

The result of FD analysis is shown in Figure-3 over a range of PESQ MOS values.

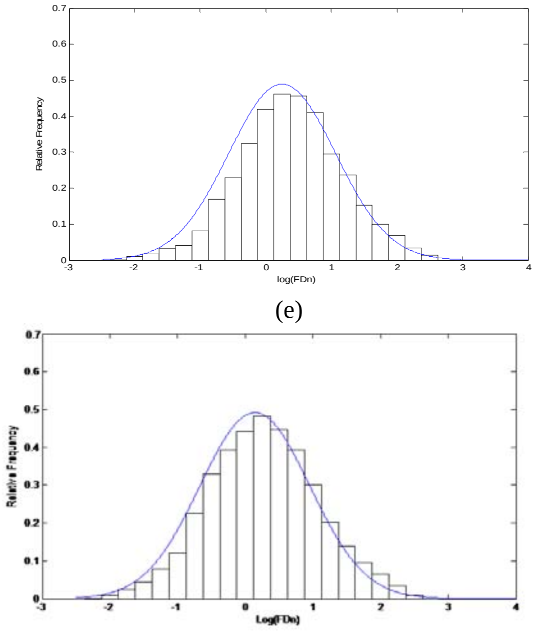


(a)

(b)

(c)

(d)



(e)

(f)

Figure-3. The $\log\left(\mathrm{FD}_n\right)$ distribution for PESQ MOS 3.0-3.5: (a) 3.0, (b) 3.1, (c) 3.2, (d) 3.3, (e) 3.4 and (f) 3.5.

The analysis of the FD shows that FD have a log-normal distribution for a given perceptual quality MOS as shown in Figure-3. The mean of distribution of $\log\left(\mathrm{FD}_n\right)$ is increasing with the degradation of the perceptual quality and vice versa as shown the Table-2. The distribution indicates that for a given perceptual quality the FD can have a wide range of values. Certain large values can be tolerated as the overall quality remains the same.

Table-2. The estimated mean, μ_0 and the standard deviation of log (FD) distribution.

PESQ MOS	Target Mean, μ_0	Standard Deviation
3.0	0.5671	0.0476
3.1	0.5340	0.0482
3.2	0.4692	0.0385
3.3	0.3559	0.0343
3.4	0.2557	0.0337
3.5	0.1550	0.0384

CONCLUSIONS

FD analysis shows that the mean of log (FD) is inversely proportional with the quality of the speech. The result of FD analysis proposes that the current practice where transmission parameters such speech codec rate adapted on a frame-by-frame is not efficient. The current practices tend to inefficient utilization of resources and possibly unsatisfactory perceptual quality. As such, to retain a required level of end-user perceptual quality, what is must is to detect the shift in the FDs distribution and then take actions to rectify that such as controlling 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. Using FER as a metric to control speech quality will result in loss of quality and/or inefficient use of radio resources. Applying this new parameter to the communication system will allow faster action at the transmitter to control the quality of the speech signals demand by the end users. The new parameter application will potentially benefit both the end users and network provider. The network provider's resources will be optimized and the end users will satisfied with the perceived speech quality.

REFERENCES

[1] K. Al-Mashoud, A. Aburas and M. Maqbul. 2012. “Speech Quality Assessment in Mobile Phone Using a Reduced-Complexity Algorithm,” in the 7th International Conference on Internet and Web Applications and Services, Stuttgart, Germany.

[2] ITU-T Recommendation P.861. 1996. “Objective Quality Measurement of Telephone Band (300-34000Hz) Speech Codec,” August.

[3] H. Hosseini, B. Rohani and B. Rohani. 2000. “Objective Characterization of Voice Service Quality in Wideband CDMA,” in IEEE VTC conference, pp. 2708-2711.

[4] W. Rix. 2004. “Perceptual Speech Quality Assessment – A review,” in IEEE Conference on Acoustic, Speech and Signal Processing, pp. 1056-1059.

[5] Featured. 2009. “Objective Perceptual Audio Quality Measurement Methods”, Broadcast Technology no. 35, Combined Issue Autumn 2008-Winter.

[6] J. B. G. Anthony W. Rix1, Michael P. Hollier1 and Andries P. Hekstra. 2001. “Perceptual Evaluation of Speech Quality (PESQ) – A New Method for Speech Quality Assessment of Telephone Network and Codecs”, in Proceeding IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP’01), pp. 749-752.

[7] Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrow Band Telephone, Networks and Speech Codecs, ITU-T Recommendation P.862, Feb 2001.

[8] A.Z Jusoh, R. Togneri, B. Rohani and S. Nordholm. 2009. “CUSUM Application in Perceptual Speech Quality Control”, in Proceeding of APPC2009, October 2009, Shanghai, China, oo. 694-698.

[9] B. Rohani, H.J. Zepernick and B. Rohani. 2004. “Application of A Perceptual Speech Quality Metric for Link Adaption in Wireless Systems,” in 1st International Symposium on Wireless Communication Systems, pp 260-264.

[10] B. Rohani and H.J. Zepernick. 2003. “Frame Erasure Pattern Feedback for Real-time Perceptual Quality Estimation,” in the 4th Pasific Rim Conference of the 4th International Conference on Information, Communication and Signal Processing, pp. 110-113 Vol.1.

[11] B. Rohani and H.J. Zepernick. 2006. “Application of a Perceptual Speech Quality Metric in Power Control of UTMS,” in 2nd ACM International Workshop on Quality of Service & Security for Wireless and Mobile Networks (Q2sWinet’06), pp. 87-94.

[12] ITU-T Recommendation. 1998. “ITU-T Codec Speech Database,” in Series P, Supplement 23, February.

[13] AMR Speech Codec Frame Structure, 3G TS 26.10, March.

[14] 3GPP TS 26.101 V6.0.0. 2004. “Mandatory Speech Codec Speech Processing Functions: Adaptive Multi-rate (AMR) Speech Codec Frame Structure (Release 6),” September.

[15] 3GPP TS 26.102 V6.0.0, “Mandatory Speech Codec Speech Processing Functions: Interface to lu, Uu and Nb (Release 6),” September 2004.

[16] B. Rohani and H.J. Zepernick. 2005. “An Efficient Method for Perceptual Evaluation of Speech Quality in UTMS,” in Proceedings International Conference on Multimedia Communications Systems Montreal, Canada, August, pp. 185-190.