IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 2, Issue 5 (July-Aug. 2012), PP 49-53
www.iosrjournals.org

Non-Intrusive Speech Quality with Different Time Scale

Mr. Mohan Singh[1], Mr. Rajesh Kumar Dubey[2]
[1] *(Dept. of Electronics and Communication, Satya College of Engineering and Technology, Palwal, India)*
[2] *(Dept. of Electronics and Communication, Jaypee Institute of Information Technology, Noida, India)*

Abstract: *Speech quality evaluation is an extremely important problem in modern communication networks. Service providers always strive to achieve a certain Quality of Service (QoS) in order to ensure customer satisfaction. Modeling the speech quality becomes an urgent issue. In this project a computable model for different time scale speech quality evaluation, called E-Model is developed. The results indicate that subjects can monitor speech quality variations very accurately with a delay of approximately 1 second. Non-intrusive speech quality is measured at the receiver from a degraded signal using G.107 (E-model) which is a parameter based model and calculate MOS values with quality rating factor. The quality rating factor is calculated by network impairments (loudness rating) of a speech. The output from the model described here is a scalar quality rating value, R, which varies directly with the overall conversational quality. The key contribution of this paper is to explore the use of G-107 (E-model) based features for different time scale non-intrusive speech quality evaluation using time varying loudness of a speech for long stimuli. Sectional speech quality is obtained by E-model, which is called instantaneous quality of the section that will be constant for each section. Overall perceived quality can be calculated by using average of instantaneous speech quality.*

Keywords: *Critical bands, E-model, loudness, MOS, non-intrusive speech quality,*

I. Introduction

The need to measure speech quality is a fundamental requirement in modern communications systems for technical, legal and commercial reasons. Speech quality measurement can be carried out using either subjective or objective methods. The Mean Opinion Score (MOS) [1] is the most widely used subjective measure of voice quality and is recommended by the ITU [2]. A MOS value is normally obtained as an average opinion of quality based on asking people to grade the quality of speech signals on a five-point scale (Excellent, Good, Fair, Poor and Bad) under controlled conditions as set out in the ITU standard [2].

In voice communication systems, MOS is the internationally accepted metric as it provides a direct link to voice quality as perceived by the end user [3]. The inherent problem in subjective MOS measurement is that it is time consuming, expensive, lack of repeatability and cannot be used for long-term or large scale voice quality monitoring in an operational network infrastructure. This has made objective methods very attractive to estimate the subjective quality for meeting the demand for voice quality measurement in communication networks. Objective measurement of voice quality in modern communication networks can be intrusive or non-intrusive. Intrusive methods are more accurate, but normally are unsuitable for monitoring live traffic because of the need for a reference data and to utilize the network. A typical intrusive method is based on the latest ITU standard, P.862, Perceptual Evaluation of Speech Quality (PESQ) Measurement Algorithm [4]. This involves comparison of the reference and the degraded speech signals to obtain a predicted listening-only one-way MOS score. Since the quality of a speech signal does not exist independently of a subject, it is a subjective measure. The most straightforward manner to estimate speech quality is to play a speech sample to a group of listeners, who are asked to rate its quality. Since subjective quality assessment is costly and time consuming, computer algorithms are often used to determine an objective quality measure that approximates the subjective rating. Speech quality has many perceptual dimensions. Commonly used dimensions are intelligibility, naturalness, loudness, listening effort, etc., while less commonly used dimensions include nasality, graveness, etc. However, the use of a multidimensional metric for quality assessment is less common than the use of a single metric, mainly as a result of cost and complexity. A single metric, such as the mean opinion score scale, gives an integral (overall) perception of an auditory event and is therefore sufficient to predict the end-user opinion of a speech communication system.

However, a single metric does not in general provide sufficient detail for system designers. The true speech quality is often referred to as conversational quality. Conversational tests usually involve communication between two people, who are questioned later about the quality aspects of the conversation; the most frequently measured quantity is listening quality. In the listening context, the speech quality is mainly affected by speech distortion due to speech codecs, background noise, and packet loss. One can also distinguish talking quality, which is mainly affected by echo associated with delay and sidetone distortion. The distorted (processed) signal or its parametric representation is always required in an assessment of speech quality. However, based on the availability of the original (unprocessed) signal, two test situations are possible: reference based and not reference based. This classification is common for both the subjective and objective evaluation of speech quality. The absolute category rating (ACR) procedure, popular in subjective tests, does not require the original signal, while in the degradation category rating (DCR) approach the original signal is needed. In objective speech quality assessment, the historically accepted terms are intrusive (with original) and non-intrusive (without original).

II. Critical Bands

The concept of critical bands is introduced in this chapter, methods for determining their characteristics are explained, and the scale of critical-band rate is developed. The definitions of critical-band level and excitation level are given and the three-dimensional excitation level versus critical-band rate versus time pattern is illustrated. The concept of critical bands was proposed by Fletcher. He assumed that the part of a noise that is effective in masking a test tone is the part of its spectrum lying near the tone. In order to gain not only relative values but also absolute values, the following additional assumption was made: masking is achieved when the power of the tone and the power of that part of the noise spectrum lying near the tone and producing the masking effects are the same; parts of the noise outside the spectrum near the test tone do not contribute to masking. Characteristic frequency bands defined in this way have a bandwidth that produces the same acoustic power in the tone and in the noise spectrum within that band when the tone is just masked. Fletcher's assumptions may be used to estimate the width of characteristic bands, and we shall see later on how these values compare with the critical bandwidths determined by other measurements.

However, the assumption that the criterion used by our hearing system to produce masked threshold is independent of the frequency of the tone is incorrect. As will be discussed later, the power of the tone at masked threshold is only about half to a quarter of that of the noise falling into the band in question. Using this additional information, the width of the bands in question, the critical bands, can be estimated quite closely. At low frequencies, critical bands show a constant width of about 100 Hz, while at frequencies above 500 Hz critical bands show a bandwidth which is about 20% of centre frequency, i.e., in this range critical bandwidth increases in proportion to frequency. In contrast with the estimation of the width of the critical band using the assumption described above, there exist several direct methods for measuring the critical band.

III. LOUDNESS

Loudness belongs to the category of intensity sensations. The stimulus sensation relation cannot be constructed from the just-noticeable intensity variations directly, but has to be obtained from results of other types of measurement such as magnitude estimation. In addition to loudness, loudness level is also important. This is not only a sensation value but belongs somewhere between sensation and physical values. Besides loudness in quiet, we often hear the loudness of partially masked sounds. This loudness occurs when a masking sound is heard in addition to the sound in question. The remaining loudness ranges between a loudness of "zero", which corresponds to the masked threshold, and the loudness of the partially masked sound is mostly much smaller than the loudness range available for unmasked sound. Partial masking can appear not only with simultaneously presented maskers but also with temporary shifted maskers. Thus the effects of partially masked loudness are both spectral and temporal.

Loudness comparisons can lead to more precise results than magnitude estimations. For this reason the loudness level measure was created to characterize the loudness sensation of any sound. It was introduced in the twenties by Barkhausen, the researcher whose name was shortened to create a unit for critical-band rate, the Bark. Loudness level of a sound is the sound pressure level of a 1-kHz tone in a plane wave and frontal incident that is as loud as the sound; its unit is "phon". Loudness level can be measured for any sound, but best known are the loudness levels for different frequencies of pure tones. Lines which connect points of equal loudness in the hearing area are often called equal-loudness contours. They have been measured in several laboratories and hold for durations longer than 500 ms. Because of the definition, all curves have to go through the sound pressure level at 1 kHz that has the same value in dB as the parameter of the curve in phon: the equal-loudness contour for 40 phon has to go through 40 dB at 1 kHz. Threshold in quiet, where the limit of loudness sensation is reached, is also an equal-loudness contour. Because threshold in quiet corresponds to 3 dB at 1 kHz and not to0 dB, this equal-loudness contour is indicated by 3 phon. Equal-loudness contours are normally drawn for a frontally-incident plane sound field. However, in many cases the sound field is not a plane sound field but similar to what is known as a diffuse sound field, in which the sound comes from all directions.

IV. Approach

For calculating speech quality with different time scale, we need to calculate loudness of that speech signal and that loudness apply to the G-107 (E-model). This whole process is given below fig. [1].

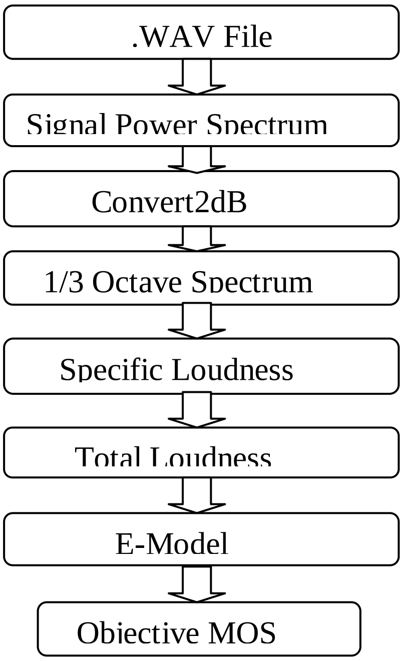

Fig.1. Calculating Objective MOS

Objective MOS is to be calculated from the above approach. First take a speech file (.wav) and calculate power spectrum of that speech file and then convert it into dB. This signal power spectrum passes through 1/3 octave filter .An octave filer is the filter whose highest frequency is double of lowest frequency which is used to find out the information of a signal, for this purpose we split the speech signal into small section that is known as bands (critical bands). 1/3 power spectrum of an octave filter is known as 1/3 power spectrum. Now for each critical band, specific loudness is to be calculated. Total loudness is the sum of all specific loudness and this loudness is to be applied to the E-model. The output of E-model is the sectional quality rating factor, R and objective MOS is calculated by using given formulae [1].

For $R < 0$: MOS = 1
For $0 < R < 100$: MOS = $1+0.035R+R(R-60)(100-R)7.10^{-6}$
For $R > 100$: MOS=4.5

V. Instantaneous Quality

Time Varying Speech Quality (TVSQ) method is motivated by the fact that the speech or voice quality of new networks varies, even during a single conversation, due to specific impairments like packet loss, handover in mobile network [5]. During a communication, the speech quality can vary due to the special technical characteristics of the different networks such as mobile or IP networks. The communication on these networks can be impaired by different factors e.g. distortion due to packet loss or bit rate, side tone, echo, compression algorithm, etc which are very common. For the assessment of speech quality, ITU-T recommended test methods are available but these methods use short speeches of ~8 seconds or ~16 seconds length. These methods are standardized and well suited for conditions when tested speech quality expected to be constant. If one wants to evaluate long speech sequence, he has to divide long speech sequences into short speech sequences of ~8 seconds or ~16 seconds length. However, if one evaluates short speech sequences then they cannot take into consideration any quality variations in time within tested speech sequence A test was conducted by France Telecom in July 1999.The purpose of the test was continuous assessment of speech quality containing quality in time and its relationship with overall subjective quality. In subjective listening test conducted by France Telecom, subjects were supposed to perform two tasks while listening to the speech sequences, instantaneous quality judgments and overall quality judgments. Instantaneous quality is the quality perceived by the subjects at any instant during the play out of the sequence whereas, overall quality is the single (scalar) judgment which subjects give after listening the whole speech sequence. After getting these two types of subjective judgments, it was analyzed whether instantaneous judgment can be used to predict overall quality. “Previous work from AT&T has shown that overall quality might be predicted by a linear contribution of short (8 sec) sound-sequences MOS (Mean Opinion Scores) assessed independently. Although the model proposed by AT&T is useful, it does not take into consideration real instantaneous judgments and hypothesizes some integration of perceived quality over 8 s. Moreover, the model was derived from stimuli that did not contain real degradations representing those found in wireless or landline packet networks”. [6]

VI. Results

Time domain speech signal of 8 sec. is given in fig. [2]. in this figure there is no voice in between 2 to 6.5 sec. Voice is present for 0-2 sec. and 6.5-8 sec

.

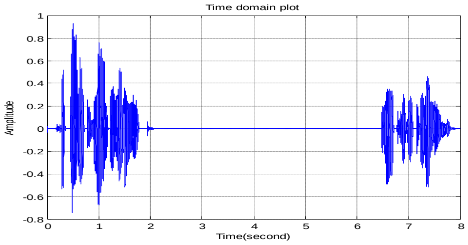

Fig.2: Time domain plot for 8 sec speech.

The 8 sec. speech instantaneous quality is given in fig. [3], in which first part is for 4 sec. and seconds part is for other 4 seconds and similarly time domain plots and instantaneous quality for 30 seconds and 60 seconds are given in fig.[4], fig. [5], fig. [6] and fig. [7] respectively.

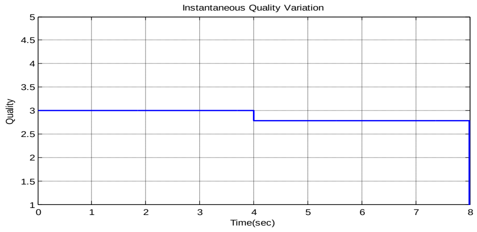

Fig.3: Instantaneous quality for 8 sec

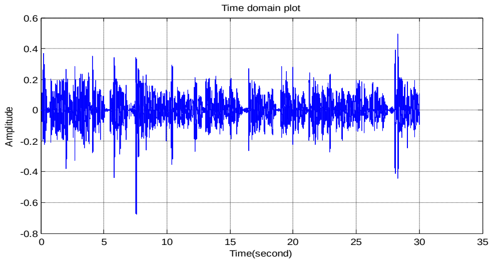

Fig.4: Time domain plot for 30 seconds speech.

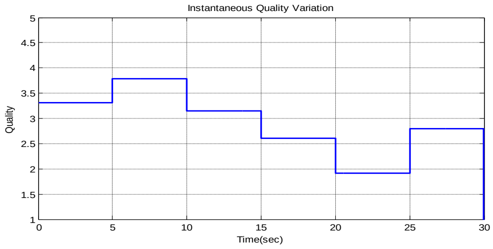

Fig.5: Instantaneous quality for 30 sec speech.

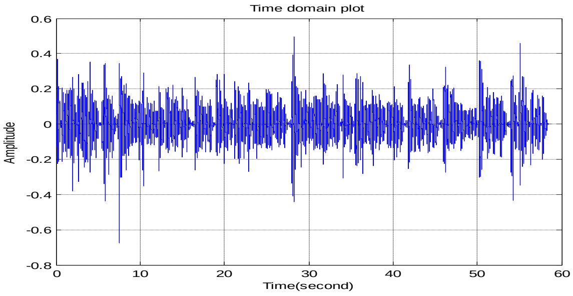

Fig.6: Time domain plot for 60 sec speech

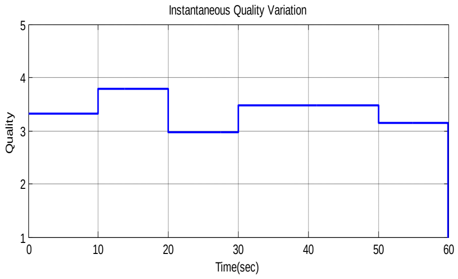

Fig.7: Instantaneous speech quality for 60 sec. speech file.

VII. Disscussion

Speech quality evaluation using e-model (ITU-T.G.107) for different time scale is calculated. Speech files are taken for 8 seconds, 20 seconds and 60 seconds respectively for sectional speech quality. This quality is obtained using loudness of this speech file by e-model which is known as non-intrusive speech quality with different time scale .The output is the sectional quality for a speech file and section quality is constant for each section. This sectional quality is different for each section due to different loudness for that section. Loudness of a speech is very important and is helpful for calculating the speech quality of a speech. Objective MOS and average objective MOS is calculated for each section for 30 sec and 60 sec speech respectively table 1.

Table 1: Sectional and average objective MOS for 30 sec and 60 sec speech

Sections	Sec-1 OMOS	Sec-2 OMOS	Sec-3 OMOS	Sec-4 OMOS	Sec-5 OMOS	Sec-6 OMOS	Average OMOS
For 30 sec	3.3169	3.7856	3.1459	2.6123	1.9200	2.7918	2.9287
For 60 sec	3.3169	3.7856	2.9702	3.4817	3.4817	3.1459	3.3637

VIII. Conclussion

Speech quality of a speech signal is obtained by conversational e-model ITU G.107 which is known as an overall speech quality .E-model is not used to find out instantaneous quality of a speech signal. In this thesis different time scale speech quality for non-intrusive is to be calculated by e-model using time varying loudness of a speech signal. Time varying speech quality is calculated for 8 sec., 20 sec. and 40 sec. speech signal. Instantaneous quality for 8 sec. speech file for sections is to be calculated and constant for each section in which first section (4 sec.) and second section (4 sec.) and finally calculated average speech quality for 8 seconds for 24 speech files. And hence time varying speech quality is depends upon time varying loudness of a speech signal and overall perceived quality can be calculated by using average of instantaneous speech quality.

References

[1] ITU-T Rec. G.107, *The E-Model, a Computational Model for Use in Transmission Planning*, 2003.

[2] ITU-T Rec. P.800, *Methods for Subjective Determination of Transmission Quality*, Int. Telecomm. Union, Aug. 1996.

[3] W. C. Hardy, "QoS Measurement and Evaluation of telecommunications Quality of Service.," John Wiley & Sons, ISBN 0-471-49957-9, 2001.

[4] 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*, Geneva, Switzerland, Feb.2001.

[5] ITU-T Rec. P.880, *Methods for objective and subjective assessment of quality*, August 2004.

[6] *Contribution ITU-T[COM 12-94]*, Continuous assessment of time-varying subjective vocal quality and its relationship with overall subjective quality, 1999.