

Perceptual Evaluation of Call Quality and Evaluation of Telecom Networks

Akram Aburas^a, Professor J. G. Gardiner^a, Prof. Khalid Al-Mashouq^b
Dr. Zeyad Al Hokail^b

^a School of Engineering, Design and Technology, University of Bradford
Bradford, West Yorkshire, United Kingdom

^b Electrical Engineering Department, King Saud University
Riyadh, Saudi Arabia

Abstract.

The drastic increasing challenges in providing high quality, reliable and low cost mobile voice services have made speech quality a major research area. With the evolution of liberalized market, the quality of voice services in telecommunication has become a significant issue. In this environment, speech quality is becoming a key factor distinguishing one operator from another and hence a strong indicator for customer satisfaction. Therefore, it is imperative that a service provider is capable of predicting customers' perceptions of quality so that networks can be optimized and maintained. The proposed method is based on determining and correlating with subjective MOS of the speech in telecommunication network. As an initiative the work is majorly concentrated in determining the call quality using the several parameters such as call attempts successful, call attempts unsuccessful, normally dropped, and dropped due to handover were used in evaluating the signal quality. We have evaluated two operators and analyzed their signal quality based on our proposed parameters.

Keywords – Speech quality, Telecommunication networks, Mean opinion score (MOS), signal strength, Symbian OS.

1. INTRODUCTION

Traditionally, user's perception of speech quality is measured offline using subjective listening tests. The method of subjective testing called Mean Opinion Score (MOS) provides a numerical indication of the perceived quality of received human speech over the connection. The MOS is expressed as a single number in the range 1 to 5, where 1 is lowest perceived quality, and 5 is the highest perceived quality. But subjective estimation by repeated listening tests at various sites within the coverage area is not practical since it is laborious, expensive and time-consuming. So, it would be much desirable to use an automatic objective evaluation system that applies a good objective speech quality measure to estimate the statistical average of subjective opinions of the typical conversational speech sentences sent through the mobile network.

Objective assessment of the speech quality in modern communication systems is typically achieved by measuring some form of distortion between the input (transmitted) and output (received) speech signals. Processing steps typically include normalization of signals power, time alignment between input and output records, perceptual modelling and determining a distance value, which is used to estimate the equivalent subjective quality score.

The results of MOS testing are expensive and impractical for testing in the field. Automatic testing algorithms were created in an attempt to formulate objective network testing similar to signal to noise ratio (SNR), bit error rate (BER) and receive signal strength (RSSI) which are used to measure the signal quality. Objective testing algorithms are also called automated quality measurement techniques. Three objective tests were developed namely:

1. Perceptual Speech Quality Measure (PSQM)
2. Perceptual Analysis Measurement System (PAMS)
3. Perceptual Evaluation of Speech Quality (PESQ)

Perception models for evaluating speech quality were jointly independently by Mike Hollier at BT Labs and John Beerends of KPN Research led subsequent innovations in the 1990s on the use of perception for voice quality assessment [1, 2]. Hollier observed that taking account not just of the amount, but also the distribution, of audible distortion could make quality predictions much more accurate. His work was taken up in 1996 by Antony Rix and forms the core of PAMS [3]. It was not until 1996, following a lengthy international study, those perceptual models

for quality assessment was first standardized. The result of this was that Beerends' model, PSQM, became an ITU-T recommendation (P.861) for assessing speech codecs [4].

Over the last decade, researchers and engineers in the field of objective measures of speech quality have developed different techniques based on various speech analysis models. Currently, the most popular techniques are those based on psychoacoustics models, referred to as perceptual domain measures [5]. In these measures, speech signals are transformed into a perceptually related domain using human auditory models. Most available objective assessment techniques are based on an input-output approach [6]. In input-output objective assessment methods, the speech quality is estimated by measuring the distortion between an "input" or a reference signal and an "output" or received signal. Using a regression technique, the distortion values are then mapped into estimated quality.

Currently there are a number of techniques that can be classified as perceptual domain measures. These include the Bark Spectral Distortion (BSD), the Perceptual Speech Quality Measure (PSQM), the Modified BSD (MBSD), the Measuring Normalizing Blocks (MNB), the PSQM+, the Telecommunication Objective Speech Quality Assessment (TOSQA), the Perceptual Analysis Measurement System (PAMS), and most recently the Perceptual Evaluation of Speech Quality (PESQ) [6], which is specified by ITU-T recommendation P.862 [7], as the international standard for testing networks and codecs. Correlation between the objective speech quality measure and the subjective quality measure is mostly used as the system (or method) performance measure. In the case of input-output based speech quality assessment, good correlations were observed, which reaches up to 99% in some cases [8].

The field of estimating the speech quality using only received speech without access to the input record is relatively new area. Most recently, a couple of attempts to develop more credible non-intrusive speech quality measurements based on perceptual analysis have been reported. In 1994, Jin Liang and R. Kubichek [9] published the first paper in the field of output-based objective speech quality using perceptually-based parameters as the speech features. Their algorithm gave some good results in special cases achieving 90% correlation. Perceptually-based methods imitate the human listening method where many parameters and environmental effects are considered. R. Kubichek and Chiyi Jin [10, 11] used the vector quantization method which yields up to 83% correlation. Vector quantization has some disadvantages that yield pessimistic results. One of these disadvantages is an inherent spectral distortion in representing the actual analysis vector. Since there is only a finite number of codebook vectors, the process of choosing the "best" representation of a given spectral vector inherently is equivalent to quantizing the vector and leads, by definition, to a certain level of quantization error. As the size of the codebook increases, the quantization error decreases. However, with any finite codebook there will always be some nonzero level of quantization error. Furthermore, the storage required for codebook vectors is often nontrivial. The larger we make the codebook, the more storage space is required for the codebook entries.

Another example of this is an output-based speech quality measure which uses only the visual effect of a spectrogram of the received speech signal, reported in [12]. A spectrogram is a two dimensional representation of time dependent frequency analysis, and contains acoustic and phonetic information of the speech signal. Framing the spectrograms into blocks and using digital image processing, the method achieved a reported correlation factor of 0.65 with the subjective score. Most recently, Gray et al [13] reported a novel use of the vocal-tract modelling techniques which enables prediction of the quality of a network degraded speech stream to be made in non-intrusive way.

A novel output-based speech quality evaluation algorithm is proposed in [14]. It is based on characterizing simultaneously the statistical properties of speech spectral density distribution in the temporal and perceptual domains. Results show that the correlations of the proposed algorithm with subjective quality scores attain 0.897 for the training data set and 0.824 for the testing data set, respectively.

In [15] a new output-based speech quality measure, which uses Bark Spectrum analysis and 5th order PLP, was introduced. The measure is based on comparing the output speech to an artificial reference signal that is appropriately selected from optimally clustered reference codebook, using the Self-Organizing Map approach coupled with an enhanced k-means technique. The average correlation of this technique reached 0.85 for Bark Spectrum and 0.61 for PLP coefficient, respectively. Both short duration of speech records and limited number of speakers count as disadvantage of this study. Only two male speakers are used one for training and the other one for testing.

In [16], Khalid Al-Mashouq and Mohammed Al-Shayee proposed a time-delay multilayer neural network model that can rate the speech quality after a proper learning stage. The learning set consists of features such as Linear Predictive Coefficients LPC and per-frame energy. A per-frame target is needed to train the neural network. This target is selected to be the Euclidian distance between the features vector of the clean and corrupted speech frames. The best correlation for speaker and text independent case reached to 0.87. However, the published result is not generalized due to limited number of speech files that had been used.

The latest research was introduced in 2004 [17]. This paper describes a new output-based approach that extends the original pseudo-reference framework by replacing the VQ codebook with a discrete hidden Markov model. Correlation between objective and subjective scores shows good performance for text dependent and text independent cases 0.88 and 0.92 respectively, but does not provide speaker independence. However, all these output-based researches are not in the mobile communication environment except in Al-Mashouq and Al-Shayee work [16].

In this work we propose a new scheme for measuring speech quality over the network from the subscriber set. This method adopts a non-traditional approach and can be used as a relative index for assessing the quality of speech. Further, the method offer a significant advantage over traditional speech quality measurement as the network operator and subscriber can evaluate the speech quality for every call made under different conditions. This can be applied in a number of ways:

1. In present day multi-operator environments, service subscribers can easily use this method to choose between different networks.
2. Telecom regulatory bodies can apply this method to enact laws that will make operators charge less for less voice quality measured from their networks.
3. Network operators can use this as a marketing tool and offer subscribers reduced tariffs whenever the speech quality of calls they make are lower than a particular threshold.

We remarked that the proposed method is part of an ongoing research that will involve investigation of metrics which are used to detect the call, obtaining the signal strength information, refreshing the value at every 5 milliseconds and finally implement the application in a symbian operating system.

2. THE PROPOSED METHOD

Our ultimate goal in this work is to develop mobile handset software which will perform an automatic speech quality assessment for every telephone call. It should give an objective score for each call along with a periodic speech quality average. This in turn will give the subscribers a practical way to assess the performance of different operators. Moreover, operators may want to give his employees a version of this software to do internal network auditing.

To achieve our objective, continuous hidden markov model (HMM)-based speech recognition system will be used to generate phonetic segmentations of the processed speech. Based on these segmentations and statistical models, different probabilistic scores will be derived for the processed speech to model human perception of the quality of communication channels. HMM is considered a basic component in speech recognition systems. The estimation of good model parameters affects the performance of the recognition process so the values of these parameters need to be estimated such that the recognition error is minimized. The model parameters are usually determined during iterative process called "training process".

One of the conventional methods that are applied in setting HMM model parameters values is Baum Welch algorithm. Baum-Welch Algorithm is one of the traditional iterative techniques that are used to estimate HMM parameters. The algorithm starts from an initial model and iteratively improves on it (updates) until convergence is reached. The algorithm is guaranteed to converge to an HMM that locally maximize the likelihood (the probability of the training data given the likelihood).

As a starting point, we have collected a cellular speech database and its original version. Then, we have to build up a robust continuous speech recognizer. Finally, we have to derive an automatic machine scores wherein its correlation with its input-output counterpart is very high. The speech databases will be used in most of the simulation stage are the well-known TIMIT and CTIMIT database set. CTIMIT is the cellular version of the TIMIT phonetic database. We choose these databases because of their enrichment of large vocabulary and mixed speakers accents. The HMMs have to be trained using clean speech database. The noisy speech database will be used for testing. Each word of the database is segmented and phonemes would be assigned. Then that word will be trained for the HMM model parameters using Baum-Welch Algorithm.

For the test word recognition, we use the degraded speech samples and use the forward or backward algorithm of HMM to find the most likelihood word in the database. The recognized word's probability is categorized on 1-5 scale of quality. Then the test scores will be correlated with the PESQ algorithm.

Our approach adapts a non-traditional approach and the aim is to use it as a relative index when assessing the quality of speech. As a step towards realization of our research, the software has been developed on symbian platform, which has been started as the signal strength measurement. This is applied in the following manner:

1. Upon a call being setup the signal level values are taken every 5milli seconds.
2. The following criterion in table 1 below is used to decide on the quality of the signal.
3. A score is then given to the sample collected every 5 milli seconds; with 5 being the best on a scale of 5; and 1 indicating very bad signal strength.
4. At the end of the call session a cumulative score is computed and based on the score (ranging from 1 to 5); the speech quality is approximately computed.

| Signal Level Range (dBm) | Classification |
|--------------------------|----------------|
| -120 to -95 | Extremely Bad |
| -95.00 to -85.00 | Bad |
| -85.00 to -75.00 | Average |
| -75.00 to -65.00 | Good |
| -65.00 to -55.00 | Very Good |

The signal strength parameters such as number of normal dropped calls, dropped due to handover(cell handover), number failed call attempts (due to network failure), successful call attempts were also analyzed for a call bundle of ten. The call quality information is plotted as landmarks to visualize the information. The below graphs shows the statistical information captured for two different networks based on our call quality parameters.

3. EMERGING RESULTS

The emerging results of the research are shown in the below graphs. The two networks of Operator-A and Operator-B were compared on our mentioned parameters. The best scores of both operators were same as shown in Figure 1. The worst scores of Operator-A and Operator-B were shown in Figure 2 and Figure 3. The call quality is measured based on the call bundle of size ten. The experiment is carried out at bundle strength of one hundred for each network over a period of six months. The call quality of bundles were recorded by the software in the log file and the information is plotted as landmarks on the map by capturing the GPS coordinates of the location during the active call.

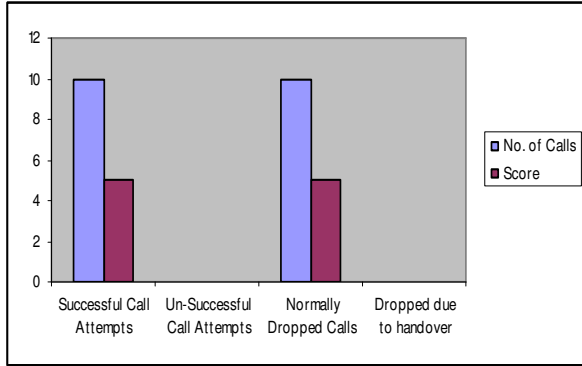


Figure 1: Operator-A and Operator-B Best Scores

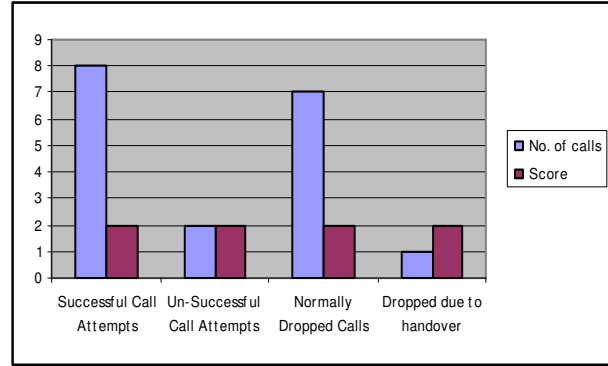


Figure 2 : Operator-A Worst Score

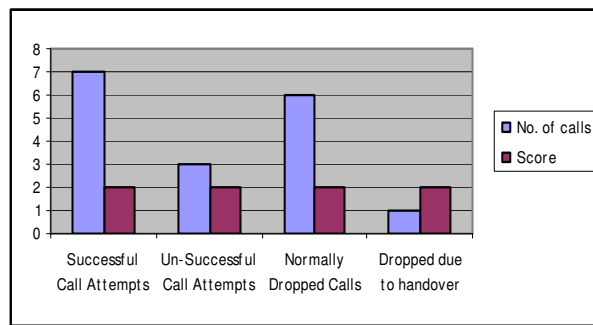


Figure 3 : Operator-B Worst Score

4. SUMMARY

This report has examined a more liberal approach to speech quality measurement in telecommunication networks. It aims to equip the network subscriber with the opportunity to choose their telecom service provider based, among other key indices, upon speech quality. This will also afford the service providers the capability of predicting customers' opinion of quality of service and the need for necessary network optimization for continued customers' satisfaction that will ensure loyalty and increased subscriber base.

5. REFERENCES

- [1] M. P. Hollier, M. O. Hawksford and D. R. Guard, "Characterisation of communications systems using a speech-like test stimulus," *Journal of the Audio Engineering Society*, 41 (12), 1008–1021, 1993.
- [2] J. G. Beerends and J. A. Stemerink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *Journal of the Audio Engineering Society*, 42 (3), 115– 123, 1994
- [3] A. W. Rix, and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, (3), 1515-1518, June 2000.
- [4] *Objective quality measurement of telephone-band (300–3400 Hz) speech codecs*. ITU-T Recommendation P.861, February 1998.
- [5] S.Voran, "Objective estimation of perceived speech quality-Part I: development of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Process.*, Vol., No. 4, p-p. 371-382, 1999.

- [6] J. Anderson, "Methods for measuring perceptual speech quality," *Agilent Technologies-White Paper*, USA, May 2001.
- [7] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), An objective method for end to end speech quality assessment of narrowband telephone networks and speech codecs," 2001.
- [8] Aruna Bayya and Marvin Vis. "Objective measure for speech quality assessment in wireless communications," *Acoustics, Speech, and Signal Processing, ICASSP-96*, IEEE International Conference 1996, vol.1, p-p. 495-498.
- [9] Jin Liang and Robert Kubichek, "Output-based Objective Speech Quality," *Vehicular Technology Conference*, 1994 IEEE 44th vol.3, p-p. 1719-1723.
- [10] Chiyi Jin and Robert Kubichek, "Vector Quantization Techniques for Output-Based Objective Speech Quality," *Acoustics, Speech, and Signal Processing, ICASSP-96*, IEEE International Conference 1996, vol.1, p-p. 491-494.
- [11] Chiyi Jin and Robert Kubichek, "Output-Based Objective Speech Quality Using Vector Quantization Techniques," *Signals, Systems and Computers, Conference Record of the 29th Asilomar Conference*, IEEE 1995, vol.2, p-p. 1291-1294.
- [12] O.C. Au and K. H. Lam "A Novel Output-Based Objective Speech Quality Measure for Wireless Communication," *IEEE Proceedings of ICSP '98*, Vol. 1, p-p. 666-669, Beijing, China, Oct. 1998.
- [13] P. Gray, M. P. Hollier and R. E. Massara, "Non-Intrusive Speech-Quality Assessment Using Vocal-Tract Models," *IEE Proc. – Vis. Image Signal Process.*, Vol. 147, No. 6, p-p. 493-501, 2000.
- [14] Chen, G. and Parsa, V. "Output-based speech quality evaluation by measuring perceptual spectral density distribution," *IEE Electronics Letters*, 40, p-p. 783-785, 2004.
- [15] D. Picovici and A.E. Mahdi, "Output-based objective speech quality measure using self-organizing map," *IEEE Proceedings of ICASSP-2003*, vol. 1, p-p. 476–479, 2003.
- [16] Khalid A. Al-Mashouq and Mohammed S. Al-Shaye, "Output-Based Speech Quality Assessment with Application to CTIMIT Database," *17th International Conference on Computers and Their Applications CATA-2002*.
- [17] Gaurav Talwar and Robert F. Kubichek, "Output-based Speech Quality Measurement Using Hidden Markov Models," *GSPx Conference*, 200

6. APPENDIX II: ABBREVIATION AND ACRONYM

| | |
|--------|---|
| BER | Bit Error Rate |
| BSD | Bark Spectral Distortion |
| CTIMIT | Cellular version of TIMIT (See TIMIT below) |
| HMM | Hidden Markov Models |
| IEEE | Institute of Electrical Electronics Engineers |
| ITU | International Telecommunication Union |
| ITU-T | ITU - Telecommunication Standardization Sector |
| MBSD | Modified BSD |
| MNB | Measuring Normalizing Blocks |
| MOS | Mean Opinion Score |
| PAMS | Perpetual Analysis Measurement System |
| PESQ | Perpetual Evaluation of Speech Quality |
| PSQM | Perpetual Speech Quality Measure |
| RSSI | Receive Signal Strength Indicator |
| SNR | Signal to Noise Ratio |
| TIMIT | A Corpus name. It is a phonetic data base that was recorded at Texas Instrument (TI) and transcribed at Massachusetts Institute of Technology (MIT), hence the name TIMIT |
| TOSQA | Telecommunication Objective Speech Quality Assessment |