Automated Speech Quality Monitoring Tool based on Perceptual Evaluation

Miroslav Voznak and Jan Rozhon
Department of Telecommunications
VSB – Technical University of Ostrava
17. listopadu 15/2172, 708 33 Ostrava Poruba
CZECH REPUBLIC
miroslav.voznak@vsb.cz, jan.rozhon@gmail.com

Abstract: - The paper deals with a speech quality monitoring tool which we have developed in accordance with PESQ (Perceptual Evaluation of Speech Quality) and is automatically running and calculating the MOS (Mean Opinion Score). Results are stored into database and used in a research project investigating how meteorological conditions influence the speech quality in a GSM network. The meteorological station, which is located in our university campus provides information about a temperature, a humidity, a dew point, a rain, a wind speed and an atmospheric pressure. The developed tool generates a call every five minutes through GSM network, a reference sample is transmitted and finally the MOS is calculated in PESQ model. The human ear performs a time-frequency transformation and in PESQ this is modelled by a short term FFT with a Hann window. The final PESQ score is a linear combination of the average disturbance value and the average asymmetrical disturbance value. The aim of our research project is to investigate a correlation between the speech quality in GSM and meteorological data. The paper describes our tool which automatically generates and calculates the MOS in GSM network.

Key-words: GSM, MOS, PESQ, Speech quality, P.862.

1 Introduction

Speech quality quality becomes an issue these days mainly because of the transition towards next generation networks, which results in increasing numbers of customers using telephone services based on the IP networks [1], [2]. Due to the inherent features and limitations of the VoIP the service providers seek the way of performing speech quality measurement and monitoring in their networks. The mentioned process gave birth to numerous algorithms suitable for performing objective speech quality measurements. Moreover some implementations of these algorithms allows for automation of the whole process of speech quality measurement for example by means of Linux scripting mechanisms.

GSM networks suffer from speech quality loss as well. Although the sophisticated algorithms are used to prevent the cumulative loss of information there are still inherent disadvantages of the networks that influence the speech quality and cause the difference in speech quality in every single call. Among these disadvantageous factors we can count the location of the user with his cell phone (open space, building…), the movement speed of the user causing the Doppler Effect to take place, or distortion from the switching between base stations to appear. In addition we need to include the effect of distortion caused by the devices working on similar frequencies to those used in GSM or signal scattering.

All the mentioned factors have been thoroughly studied and many precautions have been implemented to the GSM technology itself. Therefore the Doppler Effect does influence the call marginally, the base station switching happens so quickly, that the user does not notice the short signal loss at all and the cell phones are able to increase their signal strength to counteract the increased attenuation of the signal in the buildings.

The weather influence, however, has never been measured. In this paper we are going to describe the measuring system architecture and the basic principles we are going to use in our long term data measurement. This system is designed to measure quality of speech in the GSM networks, however, it can be modified with ease to measure this parameter in any other network. The most straightforward modification can be done by removing the GSM part and use the system as the speech quality monitoring tool in combination with some highly usable monitoring solutions such as Zabbix or Nagios.

2 State of the Art

The introductory part of this paper came with some basic knowledge from the field of speech quality measurement. The methodologies to evaluate speech quality can be sub-divided into two groups according to the approach applied, conversational and listening.

Conversational tests are based on mutual interactive communication between two subjects through the whole transmission chain of the tested communication system. These tests provide the most realistic testing environment but are they are very time consuming. Listening tests do not provide such plausibility as conversation tests but are recommended more frequently. According to methods of assessment, speech quality evaluation methodologies can be subdivided as subjective methods and objective methods. To evaluate speech quality, MOS (Mean Opinion Score) scale as defined by the ITU-T recommendation P.800 is applied [3].

In order to avoid misunderstanding and incorrect interpretation of MOS values, ITU-T published ITU-T recommendation P.800.1 [4]. This recommendation defines scales both for subjective and objective methods as well as for individual conversational and listening tests.

The subjective evaluation methods are based on evaluation by human beings (listeners), i.e. subjects. During the testing, samples are played to a sufficient number of subjects, and their results are subsequently analysed statistically. Subjects can evaluate the speech quality on a five-degree scale in accordance with the MOS model as defined by ITU-T. The best known representatives of these measurements include methods such as ACR (Absolute Category Rating) or DCR (Degradation Category Rating). Major disadvantages of these methods are high requirements on time, final evaluation being influenced by listener's subjective opinion and most of all impossibility to use them for testing in real time.

The objective evaluation methods substitute the necessity to involve humans in the testing by mathematical computational models or algorithms. Their output is again a MOS value or, depending on the algorithm applied, a different value which can be transferred to a MOS value using a suitable mapping function. The aim of objective methods is to estimate, as precisely as possible, the MOS value which would be obtained by a subjective evaluation involving sufficient number of evaluating subjects. Objective testing's exactness and efficiency is therefore a correlation of results from both subjective and objective measurements. Objective methods can be sub-divided into two groups, Intrusive and Non-intrusive.

The core of intrusive (also referred to as input-to-output) measurements is the comparison of the original sample before releasing it into a transmission chain of a communication system with the output sample, transmitted through the system (degraded).

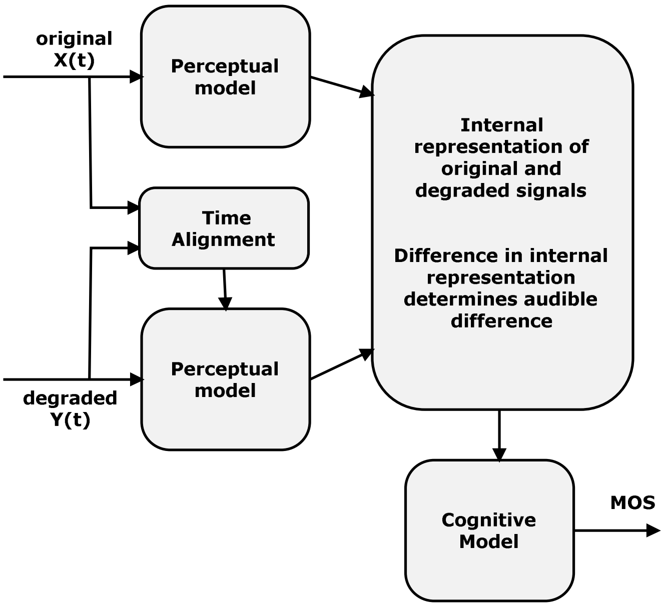
original
X(t)
Perceptual model
Time Alignment
Perceptual model
degraded
Y(t)
Internal representation of original and degraded signals
Difference in internal representation determines audible difference
Cognitive Model
MOS

Figure 1: The basic philosophy used in PESQ

This type of testing includes, among other, the following methods: PSQM (Perceptual Speech Quality Measurement), PAMS (Perceptual Analysis Measurement System) developed by British Telecommunications and PESQ (Perceptual Evaluation of Speech Quality) [5]. PESQ is the most common and most elaborate objective intrusive method. Computational technique applied by this method combines PAMS' robust temporal alignment techniques and the PSQM' exact sensual perception model. Its final version is contained in ITU-T recommendation P.862 [5], [6], [7]. The basic philosophy of PESQ approach is depicted on Fig.1, as was stated above, the princip of this intrusive test is based on comparison of original and degraded signals, their mathematical analysis using FFT and interpretation in the cognitive model.

Contrary to intrusive methods which need both the output (degraded) sample and the original sample, non-intrusive methods do not require the original sample. This is why they are more suitable to be applied in real time. Yet, since the original sample is not included, these methods frequently contain far more complex computation models. Examples of these types of measurements frequently use INMD (in-service nonintrusive measurement device) that has access to transmission channels and can collate objective information about calls in progress without disrupting them. These data are further processed using a particular method, with a MOS value as the output. The method defined by ITU-T recommendation P.563 or a more recent computation method E-model defined by ITU-T recommendation G.107 are examples of such measurements [8], [9], [12].

Today we can see various implementations of the speech quality testing mechanisms and algorithms mainly in business solutions where the companies IXIA and Sevana excel. These solutions are conformant with the specifications of the International Telecommunication Union only from some part therefore the measurements cannot be compared without the thorough knowledge of the used algorithms.

On the other hand the telecommunication union itself presents on its websites the simple implementation of one of the most advanced algorithms in the field of speech quality measurement which we have mentioned. The PESQ (Perceptual Evaluation of Speech Quality) algorithm is available for download in the form of source code and can be used to determine the conformance of the user developed solution with the ITU standard. Several open source programs are also built upon this algorithm [imankulov] and the companies (Optikom, Psytechnics), which developed the source code also offer their own services based on this source code and its modifications.

Regarding the speech quality in GSM and 3G networks mainly the first named company Optikom offers some services, but no one has performed the long term measurement with the focus on determination of weather influence on the speech quality in the GSM networks.

3 Implementation and Architecture

Our entire effort is focused on the creation of the testing platform, which would be able to gather the information in the given times and store them for a long time period. Therefore the whole solution has to be fully automated and since the Linux based systems are more straightforward in this aspect, we decided to build the testing platform upon one of the Linux distributions.

As the heart of the system we configured the Asterisk PBX, which is an open source implementation of the SIP server, to make a call regularly every 5 minutes to the SIP/GSM gateway, which is connected to Asterisk PBX with the SIP trunk over the LAN. This call is handled by the gateway and sent to the GSM network over first of two GSM interfaces equipped with the SIM card. The call is destined to the second GSM interface of the SIP/GSM gateway from where it is routed back to the Asterisk PBX.

This way we created a loop, which starts and ends on the Asterisk PBX. This approach is beneficial mainly because of the huge configuration capabilities of the Asterisk PBX.

The whole hardware equipment can be summarized in these points:

- Low-end HW Server with Ethernet interface,
- SIP/GSM gateway with two separate GSM interfaces and SIM cards.

To be able to perform the speech quality measurement using the PESQ algorithm we need two samples of speech data. The first one is the calibrated voice sample specifically created for PESQ measurements and it is used as the source file for the calls generated by Asterisk PBX. Asterisk PBX is via its dialplan rules instructed to play this file as soon as the call has been successfully established. Moreover the Asterisk “Monitor” function is used, which allow for recording both directions of the call in the separate sound files in wav container. From these two files the file containing the input channel recording is used as the second file necessary for PESQ analysis.

When the call is successfully finished asterisk calls using the “hangup” extension and “System” function the bash script, which then fires the PESQ algorithm and passes it the correct file names and arguments and after the PESQ algorithm is finished it stores the computed values into the database together with the corresponding timestamp.

The used architecture together with the most important relations among the network and software elements is depicted on the Fig. 2.

The information about current weather situation is obtained automatically and regularly in the same time interval as the speech quality measurement is performed. The nearby university meteorological station collects all the possible and measurable parameters of the current weather situation and provides them via a HTTP replies. Therefore every time the speech quality measurement is made, the bash script call the "WGET" command and retrieves the information from the meteorological station for further processing and storage.

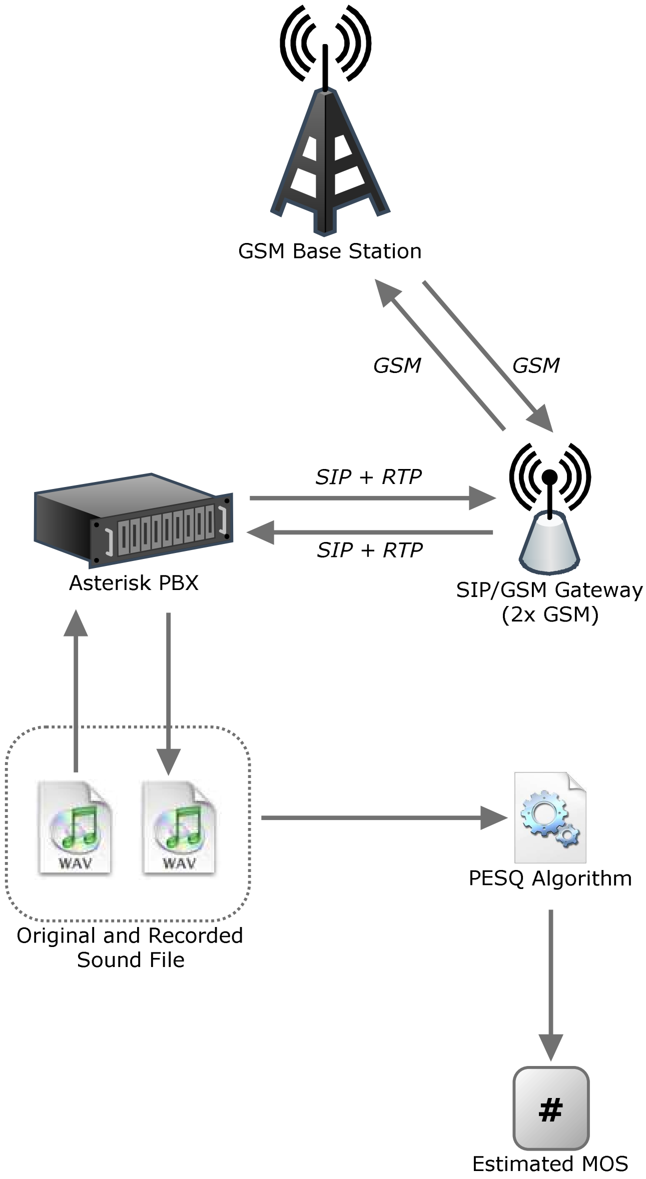

Figure 2: Testing Platform Architecture with most important relation among the network and software elements.

As a database the SQLite engine is used because of two reasons. Firstly, the configuration and manipulation of the database are quite simple in this database engine but still providing enough features to successfully complete the given task without any complication caused by the engine limitations. And secondly, the whole database can be backed up easily. This is allowed by the fact, that the whole database is stored one user-defined file. Especially the latter is important due to the long period of measurement.

Storing data into the database is useful also from the data analysis point of view, because the values of measured MOS can be displayed in almost real time with the use of the conditional database lookup for any month, day or even hour or minute thanks to the used timestamp. Moreover the aggregate functions allow quick data analysis even on the huge number of stored values.

The software equipment used to perform the measurement consists of:

- Ubuntu 10.04.1 x64,
- Asterisk PBX 1.6.2.16,
- SQLite3.

4 Basic Assumptions

From the previously presented architecture it is clear that the GSM network and the parameters affecting it is not the only factor that contributes to the final speech quality. In addition the status of the network over which Asterisk PBX and SIP/GSM gateway communicate can influence the speech quality. Moreover the codec translation between used codecs (G.711A, GSM) can have deteriorating effect on the speech quality devaluing the final results. Because these inherent effects cannot be eradicated basic assumptions applied to measurement had to be designed.

Firstly the communication between Asterisk PBX and SIP/GSM gateway takes place on the 100Mbps Ethernet line, where no other traffic is allowed. Therefore the network will never reach the congestion state and all the information on the network is exchanged as quickly as it is physically possible for this type of interconnection. Because of this the delay of packets and its variation is minimal and does not need to be counted with.

The codec translation is not a parameter, which can be easily dealt with. Since the SIP/GSM gateway supports only two codes for the VoIP communication (namely G.711 and G.729) and the GSM communication is built upon the GSM EFR codec the translation will take place every time the call is made.

Therefore the measured MOS value will always be affected by this process. To eliminate or at least diminish the influence to speech quality we do a calibration of the system. This process takes the ideal MOS value of speech quality using the GSM EFR codec as the basis, to which we compare the best result taken from several hundred measurements. The difference between the ideal and best case result is then identified as the distortion caused by the codec translation and this value is then added to all the measured values.

This way the codec translation influence is limited and the measurement will provide reasonable results. Even if this countermeasure was not performed the trend in the MOS values would still be preserved and the correlation between weather condition and measured values could still be found.

5 Results

The form of the collected data has already been outlined. The server with Asterisk PBX has all the required information at its disposal thanks to the described algorithms and scripting mechanisms. The most important information includes:

- MOS value,
- Current temperature,
- Current humidity,
- Current dew point,
- Current rain,
- Current wind speed,
- Atmospheric pressure.

Before the server performs the insertion into the database table, it also performs MOS value normalization accordingly to what was said in the previous section. After this step the insertion of data into the database takes place creating a database table as it is outlined in a simplified way in the Tab. 1.

TABLE I
Simplified Version of the Database Table

id	DATE	HOUR	MINUTE	MOS	METEOROL. DATA
1	2011-05-16	16	5	3.456	Data
2	2011-05-16	16	10	3.398	Data
3	2011-05-16	16	15	3.361	Data

In the introduction we already stated that the presented platform could also be used as a monitoring system for speech quality monitoring in VoIP systems. The basic thought is to connect the Asterisk PBX directly to the monitored line. Then the measurement could be the same as in our case.

To enhance the capabilities of the system even further, the measuring platform could be interconnected to some advanced monitoring system such as Zabbix or Nagios, which would also allow for a complex analysis including charts, or notification emails. The service could also be transformed to a commercial product that would offer the customers the possible outsourced monitoring of their VoIP connection.

Since may 2011 we have collected results more than 300 thousands measurements and we analyse them. As the convenient and efficient way of data analysis clustering algorithms allow for standardized and confirmed data segregation into the groups with similar attributes. The simplest definition is shared among all and includes one fundamental concept: the grouping together of similar data items into clusters. These clusters should reflect some mechanism at work in the domain from which instances or data points are drawn, a mechanism that causes some instances to bear a stronger resemblance to one another than they do to the remaining instances. Through several testing runs of several algorithms (K-means clustering, EM clustering…) K-means served best in our case meaning that the clusters incorporated the logically most correct data and did not suffer from the algorithm’s tendency to create the clusters of equal or similar size. We have performed particular data mining analysis and we have found out the correlations between weather attributes and speech quality and we are going to publish them in a valuable journal. In this paper, we present our developed testing tool and its relation to the research project investigating the influence of the weather conditions on the speech quality in GSM/UMTS networks.

6 Conclusion

In this paper we presented the platform for long term measurement of speech quality in the GSM network. This platform is based on the open source software allowing for automatic and regular measurement of the speech quality in the time period of several months or even more. This measurement will take place in the next five months from now with the goal to discover whether there is or is not a correlation between the weather condition and speech quality in the GSM networks.

By utilizing the key open source software such as Asterisk PBX or SQLite database, we created a robust and easily manageable system for speech quality measurement that can be used in a numerous ways. The versatility of the whole platform and the future results from the mentioned measurements are the main contribution of our work. The results will be presented after year of measurements when the whole process of collecting data finishes.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 218086.

References:

[1] I. Pravda, J. Vodrazka, “Voice quality planning for NGN including mobile networks,” in Proc. 12th International Conference on Personal Wireless Communications (PWC 2007), Prague, 2007.

[2] B. Somek, J. Herceg, M., Maletic, “Speech quality assessment, Electronics in Marine, “2004. In Proceedings Elmar 2004. 46th International Symposium, June 2004, pp. 307- 312.

[3] ITU-T P.800 – Recommendation P.800 of the International Telecommunication Union, Methods for subjective determination of transmission quality, ITU-T, 1996

[4] ITU-T P.800.1 - Mean Opinion Score (MOS) terminology, ITU-T, July 2006.

[5] ITU-T 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, February 2001

[6] Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I -Time Alignment, 2001.

[7] Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic Model, 2001.

[8] ITU-T G.107 - The E-model: A computational model for use in transmission planning, Geneva, 2009.

[9] M. Voznak, M. Tomes, Z. Vaclavikova, M. Halas, ”E-model Improvement for Speech Quality Evaluation Including Codecs Tandeming,” In Advances in Data Networks, Communications, Computers, Faro, Portugal, November 2010, pp. 119-124.

[10] M. Pravda, Z. Kocur, “Time Synchronization through Network Time Protocol and Improvement of Its Accuracy,” 32nd International Conference on Telecommunication and Signal Processing, 2009, pp. 182–186.

[11] I. Baronak, M. Halas, "Mathematical representation of VoIP connection delay," Radioengineering, Volume: 16 Issue: 3, September 2007, pp. 77-85.

[12] M. Voznak, E-model modification for case of cascade codecs arrangement, International Journal of Mathematical Models and Methods in Applied Sciences 5 (8), pp. 1439-1447, 2011.