Methodology for Voice Transmission Measurement Systems Verification

Jan Holub, Barbora Doležalová, Radislav Šmíd

Laboratory for Speech and Audio Quality Measurements, Dept. of Measurement K13138, FEE CTU Prague, Technicka 2, CZ 166 27 Prague 6, Czech Republic

tel. +420 2 2435 2131, fax: +420 2 2435 2199, holubjan@fel.cvut.cz

Abstract – The article describes methodology that has been developed for verification of measurement systems for voice transmission quality. Particular attention is given to testing of measurement algorithm implementation.

Keywords: Voice transmission quality measurements, accuracy verification

1. INTRODUCTION

Voice Transmission Quality of Service (VTQOS) measurement became an emerging task widely required by network operators, state and regulation authorities, technology vendors and system integrators. Many systems and stand-alone devices are available on the market claiming their ability for such measurements. However, it is necessary to verify their quality and accuracy since the results of such measurements influence directly serious projects of planning/extensions and innovations/upgrades.

2. STANDARDS FOR VTQOS MEASUREMENTS

There are many standards and methods for VTQOS measurement [1-3]. The mostly used intrusive method is stated in ITU-T Recommendation P.862 (Perceptual Evaluation of Speech Quality - PESQ) [3]. The measurement is performed by comparisons between original and transmitted speech sample. The speech samples have to fulfil special requirements and are usually built from several male and female voices of given language.

Except of future non-intrusive P.563 [1] the basic scheme is almost identical for all the above mentioned standards: special dedicated test call is established and suitable speech sample is transmitted between calling and called station over the tested network. Received version of speech sample is (digitally) recorded and (automatically) compared with original speech sample (after level and time alignments). This comparison is performed by algorithm that should precisely comprise all known features of human‘s ear and brain related to speech listening.

3. SPECIFIC FEATURES OF VTOQS MEASUREMENTS

The VTQOS measurement should evaluate the pair (transmitted and received speech sample) according to the above-described procedure. The result is (at least) one integral parameter, that usually corresponds to M.O.S. (Mean Opinion Score, 1=worst quality, 5=best quality). MOS can be also obtained (and really used to be obtained this way in the past) as listening tests result (according to P.830). In contrast to measurements of electrical and physical quantities, the VTQOS measurements have the following specific features:

-Inaccessibility of correct value. Usually, the conventionally correct value is obtained as results of listening tests performed on the given set of speech samples. However, it is well known that listening test results are affected by various aspects that can not be easily controlled (besides obvious ones like age and professional structure of auditorium, also some surprising factors like volume of the last bill paid for their private telephone line etc. take place). Due to this fact, even repeatability of listening tests is sometimes questionable. These problems can be partially solved by usage of trained people.

-Language and customer portfolio result’s dependency vs. demand for international and transcontinental comparisons (e.g. for interconnect agreements). As described above, the listening test results are dependent on structure of final customers. The measurement results and their accuracy are not easily transferable to different languages. On the other hand, VTQOS measurements are frequently included to interconnect agreements that can exceed not only national but even continental borders.

-Algorithm complexity. The used algorithms (like P.86x) are very computation power consuming, especially for statistical measurements performed on many (millions) of calls. Therefore, also efficiency and purity of algorithm implementation are very important and have to be evaluated.

TEST III-V		SYSTEM I			SYSTEM II		SYSTEM III			
		LQ	LE	PSQM	LQ	LE	LQ	LE	PSQM	PSQM+
F1:	Jitter30	1,36	2,29		1,25	2,19	1,768	1	6,5	14,589
	Jitter70	1	1,33	9,09	1	1,33	1,327	1	6,5	11,511
	Clip100	1	1,33	1,02	1	1,33	1,327	1,319	6,5	18,729
	Clip300	4,41	3,92	0,13	4,42	3,94	4,328	4,746	6,5	15,714
	GSMF	2,3	2,99	5,56	2,3	2,99	2,648	1,945	5,795	6,74
	GSMR	4,01	4,38	3,05	4,01	4,37	4,026	3,354	2,911	2,969
F2:	Jitter30	1,27	2,24		1,29	2,25	1,609	1	6,5	8,945
	Jitter70				1	1,48	1,327	1	6,5	9,064
	Clip100	1	1,33	0,87	1	1,33	1,327	1,11	6,5	18,513
	Clip300	3,46	2,97	0,25	3,44	2,95	3,254	3,761	0,262	1,488
	GSMF	2,4	3,16	5,25	2,41	3,16	3,009	2,232	4,693	4,922
	GSMR	3,89	4,31	2,55	3,91	4,32	4,028	3,36	2,451	2,524
M1:	Jitter30	1	1,45	6,64	1	1,53	1,327	1	5,219	5,962
	Jitter70	1	1,33		1	1,33	1,327	1	6,35	9,985
	Clip100	1,29	1,33	0,53	1,29	1,33	2,976	3,363	6,5	16,568
	Clip300	4,32	3,7	0,08	4,37	3,75	2,976	3,363	6,5	16,568
	GSMF	1,98	2,61	5,46	1,98	2,61	2,105	1,432	6,5	11,877
	GSMR	3,66	4,16	1,96	3,67	4,18	3,69	2,816	1,872	1,92
M2:	Jitter30	1	1,41		1	1,44	1,327	1	5,614	6,941
	Jitter70	1	1,33		1	1,33	1,327	1	6,5	11,533
	Clip100	1,44	1,33	0,53	1,45	1,33	2,257	2,506	6,5	16,216
	Clip300	3,81	3,39	0,16	3,83	3,4	3,978	4,276	0,191	1,61
	GSMF	1,6	2,26	5,21	1,59	2,25	1,84	1,273	5,747	7,737
	GSMR	3,82	4,28	2,11	3,83	4,29	3,192	3,153	2,007	2,058

Note: Fields marked as does not contain any data, the result there was "Correlation Timeout"

Fig.1. Example of test results. Repeatability test of c) level for 3 different tested systems

4. DEVELOPED METHODOLOGY

The suggested verification methodology has been developed with respect to above-mentioned special features of VTQOS measurements. Therefore, it differs in some points from common procedures.

-Basic parameters verification/comparison

This step covers verification of basic parameters listed in the product documentation/description and can include:

-Safety and operational requirements

-User interface features and presentation of results (graphical vs. tabular etc.)

-Hardware components evaluation

For analogue inputs (PSTN), the frequency response, AGC (Automatic Gain Control) and generating/recording device parameters – DAC/ADC (Digital-to-Analogue /Analogue-to-Digital Converter) nominal and real parameters verification, e.g. according to IEEE-STD-1241.

For digital inputs (ISDN, PCM, SDH, ATM), mainly the BER (Bit Error Rate) and FER (Frame Erasure Ratio) have to be evaluated. The basic method is short-loop measurement combined with (at least) one-direction BER measurements that enable to resolve between internal transmitter and receiver errors.

-Algorithm verification

This is the major part of the evaluation. It covers the implementation correctness verification, time consumption, accuracy and repeatability evaluation.

-Implementation correctness verification. It is performed by using known speech samples (e.g. for P.86x delivered by ITU-T). The results should fit to the prescribed M.O.S. values without differences higher than related to numerical accuracy etc. of the tested system. Time consumption is evaluated on the same set of speech samples for all the systems. The suitable speech samples sets and also the thresholds for result comparison are usually available directly in the recommendation [1], [3] in Conformance test procedure paragraphs.

-Accuracy evaluation has to be understood as most difficult and questionable step. It is necessary to use speech samples calibrated in advance by listening tests performed according P.830. All aspects discussed in Par. 3 have to be followed.

-Repeatability test of various levels is performed by

a) bringing one pair of speech samples to the system more times under different conditions (not in the same order to change initial conditions and registry content).

b) evaluation of set of pairs where the transmitted (=original) speech sample is identical and the received samples are artificially prepared, distorting the sample always by identical type and amount of impairment (e.g. white Gaussian noise at -30dB).

c) Different speech samples set deployment, where the samples are affected by the same type and amount of distortion (again, prepared usually artificially).

5. RESULTS

Three different available systems were evaluated by above-mentioned procedures. An example of measured results is shown in the Fig. 1. Due to contractual reason, only PAMS and PSQM results are shown.

Based on particular test results, also integral evaluation giving performance overview can be given.

6. CONCLUSIONS

A methodology for verification of measurement systems for voice transmission quality measurements is described. It can be used to evaluate, compare and benchmark measuring systems that are difficult to compare by standard methods.

ACKNOWLEDGEMENTS

This project is supported by the Czech ministry of Education: MSM: MSM 6840770014 ”Research in the Area of the Prospective Information and Navigation Technologies”.

REFERENCES

[1] ITU-T P.563 – Single ended method for objective speech quality assessment in narrow-band telephony applications, ITU-T, pre-published version, May 2004

[2] ITU-T P.861, Objective Quality Measurement of Telephone-band (300-3400 Hz) speech codecs, ITU-T, February 1998

[3] ITU-T P.862, Perceptual Evaluation of Speech Quality, ITU-T, February 2001

[4] Holub J.: Advanced Measurement in Mobile Networks, 4th Concertation Meeting of Mobile, Wireless and Satellite IST Projects, Brussel, March 2001

[5] Rix, A., Beerends, J.G., Hollier, M.P., Hekstra, A. P.: Perceptual Evaluation of Speech Quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, May 2001

[6] Technical documentation, applications notes and white papers for voice quality measurement systems of Agilent, Ascom, Empirix, RadCom, ECTel, Malden