Synthesized Speech Quality Evaluation Using ITU-T P.563

Ivan Kraljevski, Slavcho Chungurski, Igor Stojanovic, Sime Arsenovski

Abstract — **In this paper a method for speech quality evaluation of TTS system is presented and its usability is assessed. The ITU-T P.563 is used as a reference-free objective measurement method for speech sequences synthesized by concatenative TTS system. The method was examined and the achieved results were compared to those measured by subjective auditory tests and their correlation values were observed. It was shown that this method is useful for automatic evaluation of synthetic speech quality after major revisions of TTS systems, without the need for preparation and execution of time consuming and expensive subjective tests.**

Keywords — **ITU-T P.563, Speech quality, Text-to-Speech**

I. Introduction

SPEECH quality evaluation procedures for synthesized speech has significant role in the development and enhancement cycle of Text-To-Speech (TTS) systems. Any modification, regardless whether it is expanding the set of speech building units database (syllables, diphones, triphones etc) used for concatenative speech synthesis or improvement of the DSP algorithms used for concatenation and post-processing, will led to improvements or degradation of the TTS system performance, perceived speech quality and intelligibility.

Generally, speech quality and intelligibility are evaluated subjectively by carefully prepared auditory tests with several human listeners, which rate the recorded speech sequences by standardized procedure named MOS-Listening Quality Subjective (MOS-LQS) or Absolute Category Rating (ACR) [1]. The MOS score of a particular recording is the mean of the results reported by each of the listeners, and their values ranges from 1 - bad to 5 - excellent. Larger number of involved listeners quarantined more accurate and repeatable results. This test is standardized by the International Telecommunications Union (ITU) with the recommendation P.800, widely used in the speech and audio research and development community.

Speech quality estimation could be performed also by objective intrusive and non-intrusive measurement methods. Non-intrusive methods monitor the received speech information, where some characteristics are extracted and used for further processing for speech quality estimation. The drawback is the unavailability of the original speech sample for comparison with the distorted one and it is possible to oversee some distortion effects of the signal that are not possible to be detected or measured, but have significant influence on the perceived speech (like e.g. 3SQM P.563 ITU-T).

Intrusive methods for quality estimation use reference speech sequences that are compared with the test sequence in a similar way as the human speech perception and the quality is graded as the listeners should do in traditional subjective tests (like MOS). An example of one of the most used algorithms for intrusive tests is PESQ (Perceptual Evaluation of Speech Quality*)* [2].

For TTS quality and intelligibility evaluation, subjective auditory tests are most used. The main disadvantage is that significant resources and time is needed to perform these tests. Moreover, it is not easy to prepare and execute multiple successive experiments with same listener group due to limitations imposed by the standard regarding minimal time between two listening tests.

Because of that, it will be more convenient to use some objective measurement method to shorten the period of TTS performance evaluation between two versions after modifications or upgrades.

Several automatic instrumental measurements are known. Concatenative TTS systems with available speech corpora are characterized with quality that is inverse proportional with the number of concatenations, the quality measure could be estimated directly knowing the input text and the system, the larger the number of concatenations, the lower the quality.

Other approaches are based on the speech sequence itself and measure the spectral distance between the synthesized and reference speech sequence [3]. In case of TTS performance evaluation, it is not easy at all to create or provide reference sequences of natural speech of the same speaker from which the corpora is recorded. Even if reference sequences are available there are other issues regarding time alignment of the synthesized and reference speech sequence. This approach is useful only in case where the perceptual degradations are dependent of concatenation effects and reference sequence is available, however, practically these conditions are hard to meet.

Ivan Kraljevski, Faculty of ICT, FON University, bul. Vojvodina bb, 1000, Skopje, R. of Macedonia; (phone: +389 (2) 2445 593; fax: +389 (2) 2445 550; e-mail: ivan.kraljevski@fon.edu.mk).

Slavcho Chungurski, Faculty of ICT, FON University, bul. Vojvodina bb, 1000, Skopje, R. of Macedonia; (phone: +389 (2) 2445 555; fax: +389 (2) 2445 550; e-mail: chungurski@fon.edu.mk).

Igor Stojanovic, Faculty of CS and IT, University Goce Delcev, bul. Krste Misirkov bb, 2000 Stip, R. of Macedonia, (phone: ++389(32) 550100; fax: +389 (32) 550100 e-mail: igor.stojanovik@ugd.edu.mk).

Sime Arsenovski, Faculty for ICT, FON University, bul. Vojvodina bb, 1000, Skopje, R. of Macedonia; (phone: +389 (2) 2445 590; fax: +389 (2) 2445 550; e-mail: sime.arsenovski@fon.edu.mk).

On the other side, new standards for objective measurements like, PESQ [2] presents high correlation values with subjective auditory tests [1], other non-intrusive algorithms does not require reference speech like ITU-T P.563 standard [4].

In this paper a method for speech quality evaluation of TTS system is presented and its usability is assessed. The ITU-T P.563 is used as a reference-free objective measurement method for speech sequences synthesized by concatenative TTS system.

The used standard was developed for non-intrusive quality measurements of speech transmitted over telecommunication networks and it is optimized over speech degraded by effects of the transmission channel (noise, used codec distortions etc). So it is interesting to examine its performance compared to results achieved by subjective auditory tests. Subjective and objective measurements were performed over 3 categories of synthesized speech sequences and their correlation values were observed. The paper is structured as follows. First the measurement method is described, then the listening tests are presented, the next section introduces the usage of P.563 measure and finally the achieved results and conclusions are presented.

II. METHOD DESCRIPTION

A method for synthesized speech quality evaluation using ITU-T P.563 is presented, the speech sequences were synthesized by existing TTS system on Macedonian (TTS-MK) [5]. The system uses recorded speech corpora with female voice and it was used to create 3 distinct categories with 3 different speech sequences that were subject of auditory listening tests with large number of participants. The same set was used as well in objective measurement tests with ITU-T P.563 algorithm. Detailed analysis and comparison of the achieved results was made and the correlation values were observed form the subjective and objective measurements.

The subjective auditory listening test were prepared and performed according to modified method for TTS assessment given in [6]. For this purpose, a listening test and survey were realized aided by specially prepared WEB site, where listeners participate by listening prepared synthesized speech sequences and rate them by answering particular questions, the survey session is considered valid only if participant answers all question in consecutive order.

To ensure objectivity and validity during evaluation, several measures must be taken. Textual recourses that take part in the speech synthesis process were automatically extracted by standard methods form the textual corpora which was used as well as a basis for the TTS system. The chosen textual content must be unknown to the evaluation participants, and the synthesized speech has to be available to the each listener only once during test session, this was provided by the WEB site itself [7].

Textual materials used in the evaluation consist of three categories of speech sequences: text with meaning, semantically unpredictable sentences and a list of telephone directory.

Text with meaning category sentences were derived from existing news texts and two methods were used for their selection: selection based on a minimum word frequency and selection based on trigrams full frequency (arrays of three consecutive letters) [6]. Semantically unpredictable sentences were composed of different syntax structures randomly chosen from vocabulary of frequent short words. Telephone directory sentences list was created randomly from a telephone directory.

III. LISTENING TESTS

Participants in the auditory tests belong to the common PC user population, with the limitation that:

a) they were not directly involved in the preparation and performance evaluation of voice synthesizers or related field.
b) they weren't participants in the subjective tests, at least 6 months or similar tests within one year.
c) they do not known the content of synthesized speech sequences.

The number of the participants in the auditory tests was 150 (24 females and 126 males) aged from 13 to 79, average age 25 years. Synthesized speech sequences were stored in 16-bit linear PCM format with 16 KHz sampling frequency, with average duration of 21 sec.

The listeners have opportunity to rate the synthesized speech sequences over several characteristics: invalid pronunciation (for a word in the sentence), incorrect duration (for a word in the sentence), poor voice quality (metalized, bad concatenation etc.) poor intonation and lack of pause between words, i.e. phrases, unintelligible sentence etc.

Based on the recommendations for speech synthesizers evaluation given in [6], nine step questionnaire for the 3 categories of text (text with meaning, semantically unpredictable sentences and a list of telephone directory) with 3 sequences was prepared for the evaluation of TTS-MK synthesizer. The questionnaire was completed by all registered participants in the evaluation. For this research, only the results (scores 1-5) of these characteristics were considered: intelligibility, speech naturalness, accenting and intonation.

IV. ITU-T P.563

Recommendation ITU-T P.563 is first ITU-T standard for objective non-intrusive speech quality measurement over narrowband communications channels [4].

Speech signal which is brought to the measurement system is pre-processed before its quality is estimated. The pre-processing process consist of several phases: speech and noise signals were filtered by the modified Intermediate Reference System (IRS) filters used in ITU-T P.862 [2], speech level normalization and voiced and unvoiced parts separation with Voice Activity Detection (VAD). The next step is extraction of speech and distortion parameters from the speech segments. First the vocal tract and linear prediction coefficients are analyzed (analysis by synthesis process), where the vocal tract is represented as a series of tubes with different lengths and cross-sections. The vocal tract parameters are examined for cross-sections that indicate unnatural behaviour, also the calculated LP and cepstral coefficients were validated for values in the range of natural speech. Naturalness of the speech is assessed separately for male and female voices. In the case of robotization, additional gender independent assessment is made.

By analysing of the vocal tract and LP coefficient modification for typical human speaker, a high quality reference speech signal is synthesized. This reference signal together with the original input speech is processed with an algorithm for the baseline quality evaluation similar to ITU-T P.862 [2]. Furthermore, specific distortions, as time and amplitude clipping, interruptions and strong additional noise were detected and analyzed.

Therefore, parameters for the 6 main distortions categories is estimated: low SNR for background noise, segmental noise, interruptions and mutes in the signal, robotised speech, unnatural male and female speech.

Then, the dominant distortion class is determined and generated trough linear combination with the estimated parameters an intermediate MOS score. The final MOS score of the speech quality was obtained by a combination of the intermediate score with additional speech signal features of the speech signal. Finally, as an output, prediction for subjective ACR MOS score is presented.

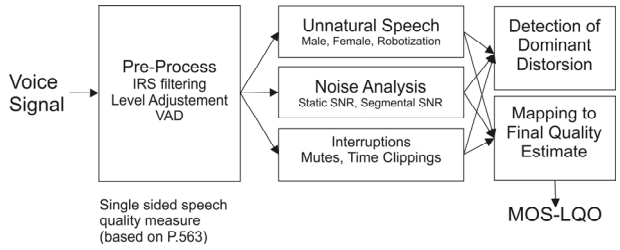

Fig 1. ITU-T P.563 speech quality measurement

For this research two experiments with objective speech quality measurement were performed, the first P.563 was used for speech quality measurement on synthesized sentences with original sampling frequency (16 KHz), even thou it is not recommended by the standard ITU-T P.563, then, the original synthesized sequences were downsampled to 8 KHz sampling frequency. In both cases, the measurement MOS values were observed. The synthesized speech quality measurement is not proposed in the ITU-T P.563 standard, nevertheless the described tests and comparisons were made to see how the algorithm will behave in the given conditions and will it be useful for synthesized speech quality assessment.

V. RESULTS

Pearson's correlations coefficients were calculated according to the achieved results of the subjective and objective tests by the given equation:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (1)$$

Table 1 shows the mean MOS scores for 3 different categories with 3 synthesized sentences per category and four different speech characteristics, obtained by subjective auditory tests, as well the general MOS scores estimated by P.563 for both cases (sampling frequency 16 KHz and 8 KHz).

Here, it could be noticed that, for the case where original synthesized sequences (16 KHz) were tested lower MOS scores were obtained compared to the case where downsampled versions (8 KHz) were tested.

However, in the first case higher values for correlation (R1) were calculated and it could be seen that the general MOS score (for P.563 16 KHz) corresponds best with the "naturalness" feature which is directly related to speech quality, that is the number and the quality of the concatenations in the synthetisation process.

In the later case very poor correlations are observed, even thou the estimated MOS scores were closer to the values of the subjective auditory tests, and knowing that female voice is implemented in the TTS system this could be explained as reported in [8] and [9].

This is due to the fact that the measurement methods based on ITU-T P.563 depends on the vocal tract analysis for assessment of the unnaturalness of the observed speech. This parameter is examined separately for male and female voice and compromise has to be made for male-female composite signals.

Considering facts that bringing speech signal with higher sampling frequency (16 KHz) as input for ITU-T P.563 measurement system, which expects signals in 16-bit linear PCM, 8 KHz, IBM format, the speech sequence effectively slows and their frequencies are moved toward lower values. The voice becomes male-like, which is favoured by the P.563 algorithm, also described in [9]. The effect of that is manifested with low general MOS scores, but with relatively good correlation to MOS-LQS tests.

TABLE 1. MOS SCORES FROM LQS AND LQO MEASUREMENTS

	1	2	3	4	5	6	7	8	9	Average	R1	R2
Intelligibility	3,5232	3,3642	3,4768	3,6225	3,5894	3,4437	3,4305	3,6093	3,7285	3,5320	0,5042	0,0323
Naturalness	3,0464	2,9801	3,0397	3,1523	3,0927	3,0464	3,1656	3,3179	3,3576	3,1332	0,7667	0,2097
Accentuation	3,1391	3,0265	3,1457	3,2450	2,9735	3,1060	3,2318	3,2450	3,3841	3,1663	0,4869	-0,1072
Intonation	3,0596	2,8874	3,0199	3,1126	3,0000	3,0066	3,0397	3,1258	3,2848	3,0596	0,6020	0,0771
Average	3,1921	3,0646	3,1705	3,2831	3,1639	3,1507	3,2169	3,3245	3,4387	**3,2228**	0,6421	0,0586
ITU-T P.563 16 KHz	2,2287	1,5232	1,3864	1,4603	2,1891	1,6054	2,5250	2,6963	2,7464	**2,0401**		
ITU-T P.563 8 KHz	3,2708	3,1342	3,1402	2,9925	3,3871	3,2967	3,3645	3,2614	3,2945	**3,2380**		

Such measurement is more useful in the process of developing a new evolution of TTS system (female voices) in order to make comparison and see if there are improvements of the speech quality, especially for particular characteristic (like naturalness).

If average values are calculated per each sentences category, for subjective and objective measurements, it could be noticed that for both cases, high correlation values are obtained for all the characteristics except "intelligibility" (Table 2).

As mentioned before already, the dependence between speech quality and intelligibility is not straightforward, P.563 based measurements estimate the speech quality, but not intelligibility, of which the lower correlation values come. Mean MOS score for all categories of synthesized speech with 8 KHz is 3,238 and is almost identical with the result achieved with subjective auditory tests 3,223, unlike the result of MOS score 2,040 measured for synthesized speech with 16 KHz.

TABLE 2. MEAN MOS SCORES OVER SENTENCE CATEGORIES

	Cat 1	Cat 2	Cat 3	Average	R1	R2
Intelligibility	3,4547	3,5519	3,5894	3,5320	0,740195	0,912126
Naturalness	3,0221	3,0971	3,2804	3,1332	0,968913	0,997883
Accentuation	3,1038	3,1082	3,2870	3,1663	0,999884	0,945939
Intonation	2,9890	3,0397	3,1501	3,0596	0,961918	0,99927
Average	3,1424	3,1992	3,3267	3,2228	0,963907	0,998961
P.563 16 KHz	1,7128	1,7516	2,6559	2,0401		
P.563 8 KHz	3,1817	3,2254	3,3068	3,2380		

Subjective votes are influenced by many factors such as the preferences of individual subjects and the context (the other conditions) of the experiment. Thus, a regression process is necessary before a direct comparison can be made. The regression must be monotonic, so that information is preserved, and it is normally used to map the objective P.563 score onto the subjective score.

Figure 1 shows the regression analysis, where the relation could be established between mean subjective and objective measurements regarding the quality of the synthetic speech over averaged scores of the three sentence categories.

VI. CONCLUSIONS

Subjective and objective speech quality measurements on synthesized speech sequences were performed over three different sentence categories. Comparison was made and their mutual correlation was investigated. The measured subjective mean MOS values from the subjective auditory tests exhibit high correlation with the MOS score estimation obtained by P.563. The examined measurement method is useful for automatic evaluation of synthetic speech quality after major revisions of TTS systems, without the need for preparation and execution of time consuming and expensive subjective tests. It should be emphasized that the method mainly concerns the quality of speech and less on the intelligibility of synthesized speech.

Although the obtained values by P.563 have high correlation, for the evaluation of TTS systems these objective measures cannot be used without linear mapping.

The proposed method described in this paper allows shortening of the time for the development of new versions of the TTS system and getting an initial impression of the general performance of the TTS system.

It allows comparison between different TTS systems in terms of quality, to locate the problematic characteristics of a TTS system, and from several versions of a TTS system to select candidates for further subjective auditory tests.

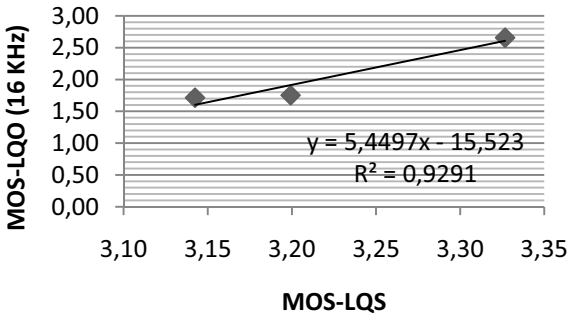

Fig 2. Regression analysis for MOS-LQS and LQO

However, objective intrusive and non-intrusive measurement methods could not replace subjective listening tests, they are able to make an estimate and clearly and quantitatively with high confidence show whether the synthesized speech has better or worse quality compared to previous versions of the examined TTS system.

VII. REFERENCES

[1] ITU-T P.800. "Methods for Objective and Subjective Assessment of Quality", August 1996.

[2] ITU-T P.862. "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to- End Speech Quality Assessment of 3.1 khz Handset Telephony (Narrow-Band) Networks and Speech Codecs", February 2001.

[3] Cernak, M., Rusko, M., "An Evaluation of Synthetic Speech Using the PESQ Measure". In: Proc. Forum Acusticum, Budapest, 2005, 2725-2728.

[4] ITU-T rec. P.563, "Perceptual Non-Intrusive Singlesided Speech Quality Measure" (2004).

[5] Chungurski, S. Kraljevski, I. Mihajlov, D. Arsenovski, S., "Concatenative Speech Synthesizers and Speech Corpus for Macedonian Language", ITI 2008, 30th International Conference on Information Technology Interfaces, 2008. 23-26 June, Pages: 669-674, Dubrovnik, Croatia

[6] Viswanathan M., Viswanathan M., "Measuring Speech Quality for TTS Systems: Development and Assessment of A Modified Mean Opinion Score (MOS) Scale", Computer Speech & Language, Elsevier B.V, Volume 19, Issue 1, January 2005

[7] S. Chungurski, I. Kraljevski, D. Mihajlov, S. Arsenovski, "Evaluation of TTS-MK System for Speech Synthesis on Macedonian Language", IX-th Nat. Conf. with International Participation, ETAI 2009, 26-29.09, Ohrid, Macedonia

[8] ITU-T Contr. COM 12-180, "Single-Ended Quality Estimation of Synthesized Speech: Analysis of The Rec. P.563 Internal Signal Processing". Federal Republic of Germany (Authors: S. Möller, T.H. Falk), ITU-T SG12 Meeting, 22-29 May, Geneva, 2008.

[9] S. Möller and T. H. Falk, "Quality Prediction for Synthesized Speech: Comparison of Approaches", Intl. Conf. on Acoustics, 2009.