Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems

Dmitry Sityaev, Katherine Knill and Tina Burrows

Speech Technology Group, Toshiba Research Europe Ltd.
Cambridge Research Laboratory, 1 Guildhall Street, Cambridge CB2 3NH, UK
{dmitry.sityaev,kate.knill,tina.burrows}@crl.toshiba.co.uk

Abstract

Evaluation of TTS systems is essential to assess performance. The ITU-T P.85 standard was introduced in 1994 to assess the overall quality of speech synthesis systems. However it has not been widely accepted or used. This paper compares the ITU test to more commonly used tests for intelligibility (semantically unpredictable sentences (SUS)) and naturalness (mean opinion score based). The aim of this research was to determine if the ITU test can provide a better performance measure and/or supplementary information to help evaluate TTS systems.
Index Terms: speech synthesis, evaluation

1. Introduction

The last decade has seen a noticeable improvement in the area of speech synthesis, especially with respect to the quality of synthetic speech. To assess quality various evaluation methods have been developed and widely used – see [1] for a good survey. As the TTS systems have improved the evaluation methods have also developed as more complex tests can be performed. However, each metric is still based on subjective assessment and the optimal metrics to use are still subject to debate.

The two major aspects of TTS system evaluation are *intelligibility* assessment and *naturalness* assessment [2]. Amongst other methods, the Semantically Unpredictable Sentences (SUS) test [3] has been used as a rigorous measure to assess intelligibility. In assessing naturalness, a Mean Opinion Scale (MOS) test [4] is probably by far the most commonly used method. These provide more detailed measures than alternative comparison-pair tests and have been recently used in large scale TTS evaluations, e.g. Blizzard Challenge, TC-STAR.

The ITU-T P.85 standard was created in 1994 with an aim to provide assessment of overall quality of speech synthesis systems [5]. The standard is based on mean opinion score (MOS) judgements across a number of different scales. There has been research into the developing and expanding the method as well as assessment of its validity and reliability [6] [7]. However, although the method has been around for more than a decade now, it has not received as wide use or acceptance as the tests previously mentioned. Alvarez and Huckvale [8] showed that the ITU test was reliable, giving similar scores across different testing sessions. However, in their experiments they found a lot of correlation across the judgement scales and this limited the ability of the ITU test to produce statistically significant differences.

This paper reconsiders the validity and usefulness of the ITU test in comparison with SUS and MOS naturalness tests. The aim of this research was to determine if the ITU test could provide a better performance measure or could provide supplementary information to complement and enhance the performance information provided by the other tests. For example, the naturalness test indicates the relative naturalness of systems compared but cannot provide detailed information as to what was the motivation for the subjective assessments by subjects. The ITU test, on the other hand, provides a series of different scores for items such as speaking rate and voice pleasantness.

2. Method

Three commercial TTS systems were used to compare the tests. The experiment was split into two sessions. In the first session, intelligibility and naturalness were assessed using the SUS and MOS. In the second session, subjects carried out the ITU test.

2.1. Test descriptions

In the SUS test, subjects were presented with sentences which had a valid syntactic structure but were semantically non-sensical (e.g. *The warm wind drank the table*) as auditory stimuli. The subjects' task was to write down each sentence the best they could. Each sentence was played only once to each subject.

In the MOS test, subjects listened to sentences from each system and had to rate the naturalness of each speech stimulus on a 10-point scale where 1 means "The utterance sounds completely unnatural" and 10 means "The utterance sounds perfectly natural". Subjects could play each stimulus as many times as they wanted.

The ITU test involved longer passages of speech – typically between 10 and 30 seconds. After the first presentation, the subjects were asked to input certain content of the message into a computer. After the second presentation, subjects rated the message using a 5-point MOS scale for each of the following: (1) Listening Effort, (2) Comprehension Problems, (3) Articulation, (4) Pronunciation, (5) Voice Pleasantness, (6) Speaking Rate, (7) Overall Quality. There was also a yes-no scale for (8) Acceptance (e.g. "Do you think this voice could be used in a commercial application such as in a car navigation system?"). Additionally, another yes-no scale was introduced called (9) Preference: "Do you personally like the voice?". A pilot study was conducted to assess different layouts for the ITU test and it was found that subjects preferred to have 6 scales at a time rather than all 9 scales. For this reason, the ITU test was

split into two sessions which differed in having either I-type scales (1)-(3) (intelligibility), or Q-type scales (4-6) (quality) alongside the other scales.

2.2. Systems

Three systems were selected for the comparisons. They will be referred to here as System A, System B and System C. They all used unit-selection techniques in order to generate the speech signal and were of similar specifications – the voice footprint was about 20MB.

A female US English voice was used in each system. (The sampling rate was 16 kHz in the case of System A and 22 kHz in the case of System B and System C). Gain normalisation was performed for each system to bring the stimuli to the same perceived loudness using the sox program.

2.3. Stimuli

Stimuli for the SUS test were created in line with recommendations in [3]. The words used in the stimuli were of sufficiently high frequency. Only monosyllabic words were used to reduce cognitive processing. To avoid overloading subjects' short-term memory, none of the sentences exceeded eight words. In total, there were 75 stimuli generated by each system (15 sentences in each syntactic category). In addition, 5 sentences (one for each syntactic structure) were constructed to be used as examples.

For the MOS test, 50 sentences were created. They were a mixture of declaratives (majority), interrogatives and imperatives. They contained one or two clauses and were all meaningful sentences of English.

The preparation of the stimuli for the ITU test was done in line with the recommendations. It was decided to select only one genre: travel directions. Each stimulus consisted of four sentences which carried the following information: route number; travel distance; exit number; road name. An example of a stimulus is: *Take Route 66. Drive for 3 miles. Take Exit 299. You are now on Presidential Boulevard.* A total of 30 stimuli were created to be used in two sessions (15 stimuli in each). 3 additional sentences were constructed to be used as examples.

2.4. Subjects

Subjects were all native speakers of American English. They were naïve as to the purpose of the experiment, and did not suffer from any hearing problems or dyslexia. They were equally balanced by gender. There were 15 subjects in the first session (SUS + MOS) and 10 subjects in the second session (ITU type Q + ITU type I). Most subjects were the same across the 2 sessions. All subjects were paid upon completion of the testing session.

2.5. Procedure

In the first session, subjects performed the SUS test and the MOS test. There were 75 stimuli presented over the headphones for the SUS test (25 from each system) and 150 stimuli for the MOS test (50 from each system). All the stimuli were randomized.

The second session which involved performing the ITU test was conducted approximately two weeks later. As mentioned above, it was split into mini experiments, involving either I-type or Q-type questionnaires. In each one, subjects were presented with 15 stimuli (5 from each system). The order of stimuli presentation was randomized.

In all the experiments, subjects were given clear instructions about how to proceed with the tests. They were also given a few practice sentences before each test. All the results were collected with the help of a computer program which was written especially for this purpose.

3. Experimental results

3.1. SUS intelligibility test

The results for the SUS tests were collected and semi-automatically processed using the HTK toolkit [9]. A sentence was considered to be transcribed correctly only if all the words (including articles) were transcribed correctly. Spelling mistakes were ignored. Homophonic substitutions were considered as correct (e.g. "weak" = "week", "sail" = "sale", etc.), irrespective of part of speech.

	System A	System B	System C
% sentences correct	52	47	56
% words correct	90	88	91

Table 1. *SUS intelligibility scores for each system.*

Table 1 shows the percentage of sentences and words transcribed correctly for each system. At both sentence and word level, the intelligibility ranking is System C > System A > System B. At word level the differences between the systems are statistically significant (System C:System B and System A:System B). The only statistical difference at sentence level was between System C and System A. Overall the intelligibility is not very high at sentence level which is not surprising since the scoring criteria are quite strict.

3.2. MOS naturalness test

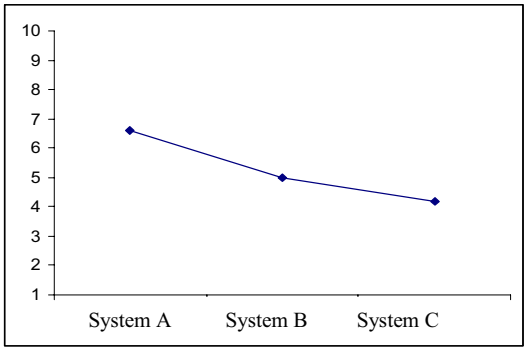

Figure 1. *Naturalness score for each system.*

Figure 1 presents the results of the naturalness assessment. The result of one-way ANOVA revealed that there was a statistical difference between the scores for the systems ($F=142.25$, $p<0.001$). Subsequent t-tests revealed that System A was more natural than System B ($df=749$, $t=1.96$, $p<0.001$), and that System B was more natural than System C ($df=749$, $t=1.96$, $p<0.001$).

3.3. ITU-T P.85 test

Table 2 presents the ITU test results for content comprehension. Since the ITU-T P.85 standard does not provide explicit recommendation on how to process this, the following approach was taken: an item was scored as correctly transcribed if the number or the name was transcribed correctly. No statistically significant differences were found between the three systems.

	System A	System B	System C
% items correct	90	92	91

Table 2. *ITU-T content comprehension scores.*

The two figures below present the MOS results grouped by I-type and Q-type scales respectively.

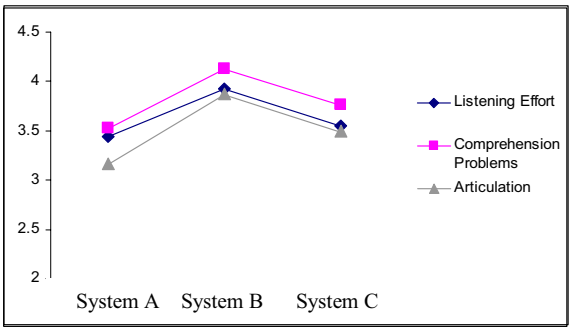

Figure 2. *ITU-T I-type MOS scores.*

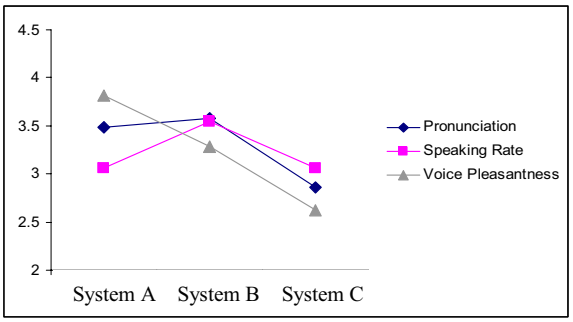

Figure 3. *ITU-T Q-type MOS scores.*

Table 3 shows the results for the Overall Quality. ANOVA analysis revealed that there was a significant difference between the three systems ($F=10.05$, $p<0.001$). System B outperformed system A which in turn outperformed System C.

	System A	System B	System C
Overall Quality	3.35	3.54	3.05

Table 3. *ITU-T Overall Quality scores.*

Table 4 presents results of the Acceptability and Preference for the TTS systems. The results displayed show that System B was found to be the most acceptable for the use in commercial applications such as in car navigation system, followed by System A and System C. However, subjects liked the voice of System A most of all and the voice of System C least.

	System A	System B	System C
Acceptability (%)	59	78	51
Preference (%)	62	42	14

Table 4. *Percentage of "yes" responses for Acceptability and Preference scales for each system.*

4. Discussion

Comparing the results for Intelligibility obtained in the SUS test and in the content comprehension ITU test, it can be seen that both tests show that word/item comprehension rate is quite high – about 90%. However, on the ITU test, no statistical significance was discovered between the three systems. The SUS test, on the other hand, revealed that System C was often more intelligible than the other two systems. This could be due to the difference in the stimuli. The stimuli in the ITU test were mostly made up of numerals and proper names. Numerals were recognised correctly more often than street names. Although some mistakes in recognition were similar between the SUS and ITU test (e.g. consonant confusion), it was felt the SUS test provided more rigorous and informative data about mis-recognitions. In addition, the system ordering from the I-type ITU questions shows System B > System C, contradicting the SUS results. Unlike the SUS and content comprehension tests, the I-type scores are subjective and it was reported in [8] that they are correlated with other effects such as acceptability. Overall, for intelligibility testing these results imply that the SUS test is generally more useful. The ITU test is probably best suited to testing intelligibility of application specific vocabulary.

As in [8], a relatively high correlation was observed over the rating scales belonging to the I-type questionnaire: Listening Effort and Comprehension Problems ($r=0.73$); Comprehension Problems and Articulation ($r=0.75$); Listening Effort and Articulation ($r=0.76$). This is unsurprising, since the questions asked for these two scales very much complement each other - if the subjects found certain words hard to understand, the more effort it could require them to understand the message. This suggests that these 3 questions could be collapsed into a single question.

The naturalness results obtained from a 10-point MOS naturalness test put System A (6.0) significantly above System B (4.5) (and System C which scored lowest). From the ITU test the Overall Quality scale would be expected to be most closely correlated with naturalness, since it is assumed to be a measure of how close the subjects believe the synthesised speech signal is to the ideal i.e. real speech. However in the ITU test, System B received a higher score for overall quality than System A and System C. It may be the case that assessing a system at different levels of prosodic structure (sentence vs. paragraph) may highlight different problem areas within the systems, which may in turn influence the subjective assessment. For example, for shorter stimuli the voice pleasantness may dominate the naturalness assessment (System A scored higher on this than the other systems) whereas other factors such as speaking rate may become more noticeable on longer stimuli.

An interesting point emerges when the results for Acceptability ("Do you think this voice is acceptable for commercial applications such as in car navigation systems?") and Preference ("Do you personally like the voice?") are compared. Whilst the subjects found System B by far the most acceptable (78% said yes), it is the voice of System A that subjects liked most. So Acceptability does not necessarily arise just from the speaker's personal preference for the voice. Acceptability may be a result of various factors, which in turn contribute to the Overall Quality rating; Acceptability, quite possibly, may just mean that a system is of sufficiently high standard to be deployed in an application. However, when it comes to selecting from several systems of comparable quality, subjective preferences will play more of a role in the decision making process, i.e. whether people like the voice or not.

At this point, it is unclear what exactly contributes to a person's choice of whether he/she likes the voice. The preference for a voice seems to correlate with the ratings on the Voice Pleasantness scale (System A > System B > System C), however this may not be the only factor. It appeared that System C had occasional flaws due to signal processing and this may have contributed to a lower score on Voice Pleasantness and a low level of liking. Whether subjects' choice is purely determined by a subjective preference for this or that voice talents' voice qualities or whether other factors are also involved is a subject for further research.

In [8] all the ITU scales were found to be correlated. Clearly this was not the case in this experiment for the ITU-Q scales. As synthesized voices become more human-like, subjects are probably more able to concentrate on subtle differences between TTS systems, such as voice quality for example (cf. Voice Pleasantness). In the past, a system with significantly bad signal processing would probably score very low on most scales compared to a system with good signal processing.

Finally, a striking observation is the relationship between the results from the SUS and the MOS Naturalness tests. System C scored highest for intelligibility yet received the lowest score for naturalness (and Overall Quality). Similar mismatches between Intelligibility and Acceptability have been previously reported. For example, [10] found that there was no strong correlation between Intelligibility and Acceptability: systems which scored highest on Intelligibility received lowest scores for General Quality.

5. Conclusions

The best approach to use to evaluate TTS systems is still an open and evolving question. Typically synthesised speech is assessed in terms of its intelligibility and naturalness. Tests such as the Semantically Unpredictable Sentences test and a MOS test, respectively, are commonly used. This study attempted to investigate whether the ITU-T P.85 standard could provide a better evaluation approach or supplementary information to these standard tests to improve the evaluation process.

Whilst the results for intelligibility were comparable in terms of a high percentage of correct recognitions, the SUS test was found to provide a more rigorous measure of which systems were more intelligible than others. Similar trends in mis-recognitions were identified between the two tests. However, the SUS test revealed more errors which could be grouped. Overall the ITU test is probably most useful for testing intelligibility of specific application items rather than as a general purpose test.

With respect to naturalness, the ITU scales revealed a different picture to that of the MOS naturalness test. For example, System A, the most natural voice by the MOS test, was rated as the most pleasant and preferred voice but System B was ranked as having the best overall voice quality. These results suggest that the ITU-Q and general questions can provide more fine grain information about the performance of a system than the single MOS naturalness test. Since the ITU test is run over longer stimuli this may also be highlighting effects that are not seen on short stimuli.

Overall, the ITU-T test can be seen to be somewhat flawed. It is not suitable for general intelligibility testing and a number of the scales are highly correlated. However, it does provide additional and complementary information with respect to naturalness and voice acceptability. With the growth in more expressive systems, more detailed assessment of these qualities is likely to become more important.

6. References

[1] Gibbon, D., Moore, R., and Winski, R. Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, Berlin, 1997.

[2] Kamm, C., Walker, M., and Rabiner, L. "The role of speech processing in human-computer intelligent communication", Speech Communication, Vol. 23, 1997, pp 263-278.

[3] Benoit, C., Grice, M. and Hazan, V. "The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences", Speech Communication, Vol. 18, 1996, pp 381-392.

[4] CCITT "Absolute category rating (ACR) method for subjective testing of digital processors", Red Book, Vol. V (Annex A to Suppl. 14), 1984.

[5] ITU-T Recommendation P.85, "A method for subjective performance assessment of the quality of speech output devices", International Telecommunications Union publication, 1994.

[6] Polkosky, M. and Lewis, J. "Expanding the MOS: development and psychometric evaluation of the MOR-R and MOS-X", International Journal of Speech Technology, 6, 2003, pp 161-182.

[7] Viswanathan, M., and Viswanathan, M. "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale", Computer, Speech and Language, Vol. 19, 2005, pp 55-83.

[8] Alvarez, Y. and Huckvale, M. "The reliability of the P.85 standard for the evaluation of text-to-speech systems", In Proc. of ICSLP, 2002.

[9] Young, S.J., Kershaw D., Moore G., Odell J., Ollason D., Valtchev V. and Woodland P.C., The HTK Book, Cambridge University Engineering Department, URL: http://htk.eng.cam.ac.uk/, 2002.

[10] Sluijter, A., Bosgoed, E., Kerkhoff, J., Meier, E., Rietveld, T., Sanderman, A., Swerts., M. and Terken, J. "Evaluation of speech synthesis systems for Dutch in telecommunication applications", In Proc. of 3rd ESCA Workshop on Speech Synthesis, 1998.