THE RELIABILITY OF THE ITU-T P.85 STANDARD FOR THE EVALUATION OF TEXT-TO-SPEECH SYSTEMS

Yolanda Vazquez Alvarez & Mark Huckvale

Department of Phonetics and Linguistics
University College London, U.K.
uva_es@yahoo.com, M.Huckvale@ucl.ac.uk

ABSTRACT

An evaluation of the reliability of the ITU-T P.85 recommended standard for the evaluation of voice output systems was conducted using six English TTS systems. The P.85 standard is based on mean-opinion-score judgements of a listening panel on a number of rating scales. The study looked at how the ranking of the six systems on the scales varied across four different text genres and across two listening sessions. Rankings were also compared with a much simpler pair-comparison test across genres and listening sessions. For the ITU test a large degree of correlation was found across scales, implying that these were not really testing different aspects of the systems. There were surprisingly similar results across sessions, implying that listeners were indeed making real judgements. In comparison, the pair comparison test gave (almost) identical rankings for systems with far less variability, making statistically significant comparisons between systems possible, even across genres.

1. INTRODUCTION

Although the ITU-T Recommendation P.85 "A Method for subjective performance assessment of the quality of speech output devices" [1] is perhaps the only published standard for speech synthesis evaluation, it has not received wide use or acceptance. This may be due to the perceived complexity of the test, or of its reliance on rating scales. Sluijter et al [2] question the validity of using five-point rating scales to evaluate the quality of a signal in the absence of either a real task or a reference natural voice. Johnston [4] questions whether the use of quality rating makes sense for signals with non-natural forms of distortion – not channel effects, but fluctuating distortions in spectrum and time. He suggests comparisons should be made to natural speech warped in frequency and time. These criticisms may be unimportant if it can be shown that a test like ITU-T P.85 can deliver consistent results with small enough variability to be able to usefully compare one system with another.

In this study we performed an ITU style test and looked at the consistency of subjects' judgements across two listening sessions held a week apart. We also looked at the selectivity of the test to differentiate between systems and between genres, as compared with a listening test based simply on direct pairwise comparisons.

2. METHOD

The study evaluated synthetic speech from six commercial text-to-speech systems for English. Two types of evaluation were used: one based on the ITU standard, and one based on pairwise comparisons. Four different generes of material were evaluated and the whole evaluation was performed twice to check retest reliability.

2.1 Evaluation Methods

In the ITU test, listeners are presented with speech from a single system and are asked to rate the signal on the basis of several scales. Different, but comparable, utterances are used for each system within one session. There are eight rating scales proposed: Overall impression, Listening effort, Comprehension, Articulation, Pronunciation, Voice Pleasantness, Speaking Rate, and Overall acceptance. The first six scales use five-point scales (shown in Table 1) in which '1' means poor and '5' means good. The Speaking rate scale uses a five-point scale in which '1' is too slow and '5' is too fast. The Overall acceptability scale uses only a two-point scale. Since these last two scales operated in a different manner, we chose not to use them in our study. We felt that our listeners would find it easier to make uniform types of judgement.

The ITU test was compared to a preference test based on direct comparisons. In the Pair Comparison (PC) test, listeners are presented with the same sentence produced by two different systems and are asked to indicate which one they 'prefer'. This type of test has been used to test system overall acceptance [3] and to determine the preference ranking of speech produced under different conditions [4].

2.2 Systems

All six systems were modern text-to-speech systems employing concatenative signal generation techniques. The systems were: Lernout & Hauspie RealSpeak (RS), Lucent technology (LT), Speechworks Speechify (SW), Elan Informatique (EL), AT&T Next Generation (ATT) and Aculab (AK). Speech materials for the first five systems were generated using their web interface; while for the last system, materials were kindly processed for us by Aculab Ltd. All materials were collected in June 2001. A female English voice was used for every system, and stimuli were produced with the default settings defined by each interface. The sampling rate was set to telephone quality when possible, but in any case all signals were subsequently band-pass filtered between 300 and 3500Hz

Rating scales	Ratings
Overall impression: "How do you rate the overall quality of the sound?"	1: Bad 2: Poor 3: Fair 4: Good 5: Excellent
Listening effort: "How would you describe the effort you needed to understand the message?"	1: No meaning understood with any feasible effort 2: Effort required 3: Moderate effort required 4: Attention necessary; no appreciable effort required 5: Complete relaxation possible; no effort required
Comprehension problems: "Did you find certain words hard to understand?"	1: All the time 2: No,not very clear 3: Fairly clear 4: Yes, clear enough 5: Yes, very clear
Articulation: "Were the sounds distinguishable?"	1: No, not at all 2: No, not very clear 3: Fairly clear 4: Yes, clear enough 5: Yes, very clear
Pronunciation: "Did you notice any anomalies in the pronunciation? "	1: Yes, very annoying 2: Yes, annoying 3: Yes, slightly annoying 4: Yes, but not annoying 5: No
Voice Pleasantness: "How would you describe the voice?"	1: Very unpleasant 2: Unpleasant 3: Fair 4: Pleasant 5: Very pleasant

Table 1. ITU-T P.85 Rating Scales used in the experiment

(telephone quality) and downsampled to 11025 samples/s. All recordings were also changed to approximately the same perceived loudness by setting the RMS level of the signals to the same absolute value (-18dB re: a maximum amplitude sinusoid in 16-bit samples). Possible differences in mean pitch or speaking rate were not changed.

2.3 Materials

Each system was aked to produce every utterance; and the same input text was used with each system. The two evaluation methods used the same materials Materials were prepared in four different genres: e-mail (em), news (n), catalogue entries (ce) and name-address-phone numbers (nap). For each genre, six utterances were designed according to the ITU-T recommendation so that different sentences were of comparable difficulty and complexity. Since some web interfaces has limits on the amount of text that could be enetered, the materials were kept to a maximum of 30 words or 160 characters. A total of 24 audio files were prepared for each system, and each lasted between about 10 and 15 seconds. This is at the lower end of the ITU duration recommendation.

2.4 Subjects

A total of thirty-six different listeners were used in the experiment. Their ages ranged from 16 to 50 years old. All of them had English as their native and dominant language (26 British English, 5 American, 1 Australian, 1 Canadian, 3 South African). None had any known hearing impairment. They were not paid for their participation. Eighteen listeners performed the ITU test and eighteen listeners performed the the paired comparison test.

2.5 Procedure

Subjects were previously informed about the aim of the experiment itself and that it had to do with *telephone-based applications* for computer voice output. A computer program SoundJudge [5] was written for this experiment. The program controlled the presentation of the audio files, presented the listener with the rating scales and logged all responses. The program allowed the listener to take the test at their own pace and to rest whenever required.

Since the ITU test used six rating scales, the subject heard each utterance twice: once to respond to scales 1-3 and once to respond to scales 4-6. On the ITU test each subject heard 24 different utterances, but the selected utterances were changed from listener to listener to ensure all systems were evaluated on all utterances. The order of presentation of systems was also randomised across listeners. Each subject made a total of 144 judgements in each session. A total of 18 judgements per system per genre per rating scale were collected.

In the pair comparison test, subjects heard a pair of identical utterances from two different systems and judged which was most preferred. They were allowed to hear the samples again if they wanted. Listeners were not able to judge the samples as equivalent. The 720 possible pairs of utterances were divided among blocks of six subjects, so that each subject only made 120 judgements. A total of 2160 judgements were collected.

A retest was conducted a week after the first test. Each subject performed the same task, with the same materials, presented in a different random order.

3. RESULTS

3.1 ITU-T test

Univariate General Linear Model Analysis (UNIANOVA) was carried out on the whole data gathered from the ITU-T test. This analysis was complemented by Scheffe and Tuckey's Honestly Significant Difference (HSD) post-hoc tests in order to find which systems differed in the subjective scores.

Table 2 shows the MOS for the different systems across the six scales. A significant effect of system was found for all scales but a large degree of correlation was found across scales. The overall ranking was ATT>SW>RS>AK>LT>EL. There was no evidence from the homogeneous subsets that any one system was ranked differently on any single scale. The ATT and SW systems were never separated, which makes sense given that they share a lot of technology.

	Sig	AK	ATT	EL	LT	RS	SW	**Homogeneouous sets**
Overall impression	*	2,28	3,74	1,73	2,19	3,13	3,60	(ATT+SW)>(RS) >(AK+LT)>(EL)
Listening effort	*	2,83	3,72	2,28	2,81	3,43	3,70	(ATT+SW+RS)> (AK+LT)>(EL)
Comprehension problems	*	3,02	3,91	2,47	2,97	3,63	3,88	(ATT+SW+RS)> (AK+LT)>(EL)
Articulation	*	2,67	3,81	2,21	2,65	3,34	3,69	(ATT+SW)>(RS) >(AK+LT)>(EL)
Pronunciation	*	2,60	3,30	2,04	2,60	3,17	3,49	(ATT+SW+RS)> (AK+LT)>(EL)
Voice pleasantness	*	2,49	3,65	2,01	2,10	3,05	3,60	(ATT+SW)>(RS) >(AK)>(LT+EL)

Table 2. Summary of the results of UNIANOVA carried out on each of the six five-point subjective scales. 'Homogeneous sets' are sets of levels of the factor 'system' which do not differ significantly.

Table 3 contains MOS for the different genres for the Overall impression scale. UNIANOVA showed that the effect of genre was significant ($p<0.001$), in particular performance on the 'nap' genre was significantly different to the others. All systems apart from AK were rated more poorly on this genre. However there is no evidence from homogeneous subsets that any one system was ranked differently on any single genre.

SYSTEM	ce	em	n	nap
AK	2,25	2,64	2,11	2,14
ATT	4,14	3,75	3,75	3,31
EL	1,67	1,91	1,83	1,51
LT	2,06	2,25	2,47	1,97
RS	3,22	3,14	3,22	2,94
SW	3,86	3,53	3,61	3,42

Table 3. Genre MOS per system for Overall impression scale

Table 4 shows the MOS for the two different listening sessions for the six scales across all systems. UNIANOVA showed that the effect of session was significant for the 'Listening effort' ($p=0.009$) and the 'Comprehension problems' ($p=0.001$) scales only. Generally there was a high degree of consistency in the mean rating and in the system ranking.

SESSION	OI	LE	CP	A	P	VP
Session 1	2,76	3,06	3,22	3,04	2,83	2,78
Session 2	2,80	3,20	3,41	3,08	2,94	2,86

Table 4. Session MOS over all scales

3.2 Pair comparison test

The pair comparison test results were calculated from the total number of preference judgements made for each system across all listeners. The use of preferences artificially highlights differences between systems.

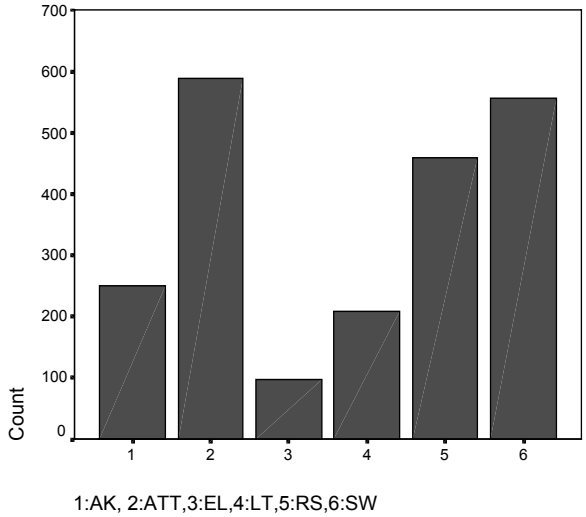

Figure 1. Overall preferences per system

Figure 1 shows the preference counts across systems for all genres. The overall ranking was the same as the ITU test: ATT>SW>RS>AK>LT>EL. To test for the significance of differences, a two-related samples sign test was used between adjacent systems in the rankings. This showed that systems ATT and SW were equally preferred, but that all other adjacent pairs were significantly different. This is a stronger outcome than was achievable with the ITU test, where systems were grouped into homogeneous subsets.

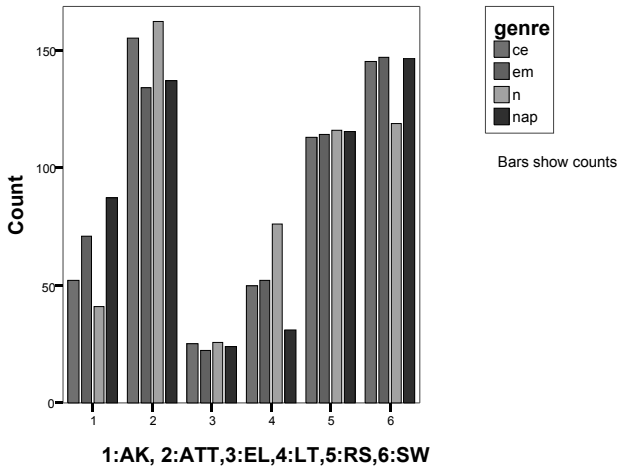

Figure 2. Overall preferences per genre

Figure 2 shows the preference scores across genres. Here it is possible to see minor variations across systems. For example a sign test shows significant differences between ATT and SW when dealing with 'n' and 'nap' genres, although these could not be told apart in the ITU test.

The rankings across genre obtained in the PC test were (almost) identical to the ones obtained from the ITU-T test. The only difference being in the 'em' genre where the ranking of ATT and SW changed places.

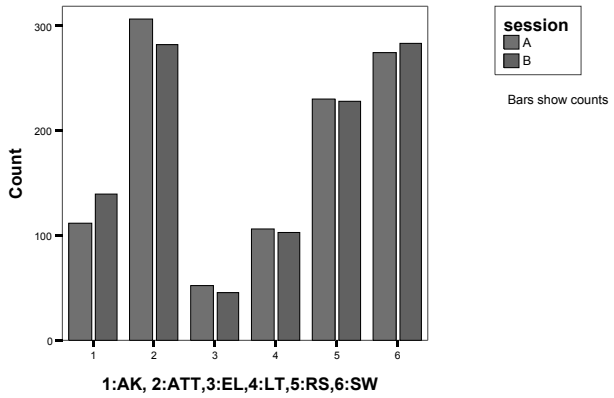

Figure 3. Overall preferences per session

Figure 3 shows the preferences across systems for the two different sessions. Again, as it was the case with the ITU-T test, subjects' preferences were very similar from one session to another. The only significant shift in preference across sessions were for systems ATT and AK (Z= -2,462; p=.014 / Z= -2,707; p=.007); however given that the total number of preferences was fixed, these may not be independent effects.

4. CONCLUSIONS

The goal of the experiment was not to determine which system was best, but to evaluate the reliability of the tests themselves and their ability to differentiate between systems and genres.

In terms of reliability both the ITU and PC tests gave very similar results when undertaken twice by listeners a week apart. This reliability was shown in terms of absolute scores and in terms of ranking. However it was also shown that the different rating scales of the ITU test were used almost identically by listeners – there was little evidence that the listeners were using these scales differently for the different systems. This calls into question whether there is much to be gained from using these.

In terms of selectivity, both the ITU and PC tests showed minor differences between the systems across genres, but only in the PC test did these differences reach statistical significance. This was the case even though we used a fairly conservative non-parametric statistical test for differences in preference. Significance may have been reached if we had used more subjects for the ITU test, but we set up the two tests to use approximately the same total number of listener judgements.

Overall we were surprised how well our listeners coped with the demands of the ITU test, and how much their judgements could be relied upon.

5. REFERENCES

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

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

[3] Kraft V., Portele T., "Quality of five German Speech Synthesis Systems. Acta Acustica 3, 351-365, 1995.

[4] Johnston, R.D., "Beyond intelligibility – the performance of text-to-speech synthesizers", BT Technology Journal, 14, 100-111, 1996.

[5] Soundjudge v1.0, Mark Huckvale, UCL, London, 2001. http://www.phon.ucl.ac.uk/resource/sfs/.