Statistical analysis of the Blizzard Challenge 2007 listening test results

Robert A. J. Clark, Monika Podsiadło, Mark Fraser, Catherine Mayo, Simon King

Centre for Speech Technology Research, University of Edinburgh
robert@cstr.ed.ac.uk, monika.podsiadlo@gmail.com, m.e.fraser@ed.ac.uk,
catherin@ling.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

Blizzard 2007 is the third Blizzard Challenge, in which participants build voices from a common dataset. A large listening test is conducted which allows comparison of systems in terms of naturalness and intelligibility. New sections were added to the listening test for 2007 to test the perceived similarity of the speaker's identity between natural and synthetic speech. In this paper, we present the results of the listening test and the subsequent statistical analysis

Index Terms: Blizzard Challenge, speech synthesis, evaluation, listening test

1. Introduction

For a full description of the 2007 Blizzard Challenge, see [1]. In the current paper, we present and discus the statistical analysis of the listening test results. Much of this analysis was made available to participants to use when publishing descriptions of their systems (see other papers in this volume).

Each participant submitted up to three voices (A: full data set of about 8 hours speech; B: ARTIC subset of about 1 hour; C: participant-selected subset of about 1 hour), and each voice was evaluated in a listening test with 5 sections. The nature of the different task in each section leads to differences in the way in which the results are presented and interpreted. There were essentially three different types of task presented to subjects to perform: Lickert-type scale rating tasks, pairwise comparisons, and type-in tasks.

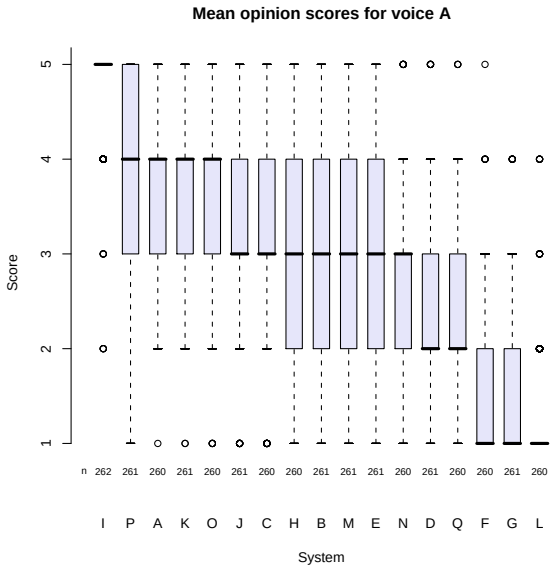

Figure 1: Boxplot showing similarity scores between systems and the original speaker for voice A, which was built from the full data set. System I represents natural speech.

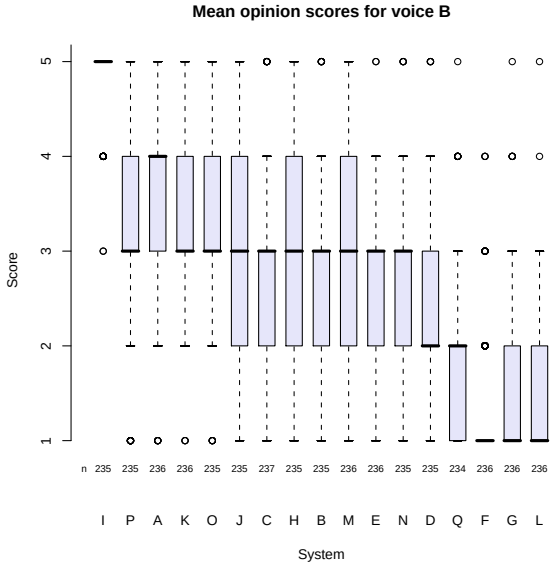

Figure 2: Boxplot showing similarity scores between systems and the original speaker for voice B, which was built from the ARCTIC data set. System I represents natural speech.

1.1. Likert-type scales

The first type of task, used in sections 1, 3 and 4 of the evaluation, asked subjects to rate individual utterances using a *Likert*-type psychometric response scale [2]. Five point scales were used in rating the similarity of a stimulus to the original target speaker in section 1, where the scale end-points were labelled "1 - Sounds like a totally different person" and "5 - Sounds like exactly the same person" and in sections 3 and 4 to determine Mean Opinion Scores (MOS) for utterances in the conversational and news domains respectively; here the scale endpoints were labelled "1 - Completely Unnatural" and "5 - Completely Natural". The internal points of the scale also received appropriate labels.

Likert-type scales are inherently **ordinal** scales. That is, the points on the scale have a ordering but there is no guarantee that the interval spacing between points is equal. A consequence of this is that it is not appropriate to compare means of judgements on such scales without first determining that the scales are behaving as intervals. This could be done, for example, by a direct comparison to rating of the same data on a known ordinal scale, such as that obtained by magnitude-estimation or a similar technique. Note that it **is** valid to calculate means for this data, but it is **not** statistically meaningful to compare them. In other words, one cannot say that the mean for one system is significantly different from any other. See [3] for a detailed summary of the specific issues that arise when treating ordinal scales as interval scales.

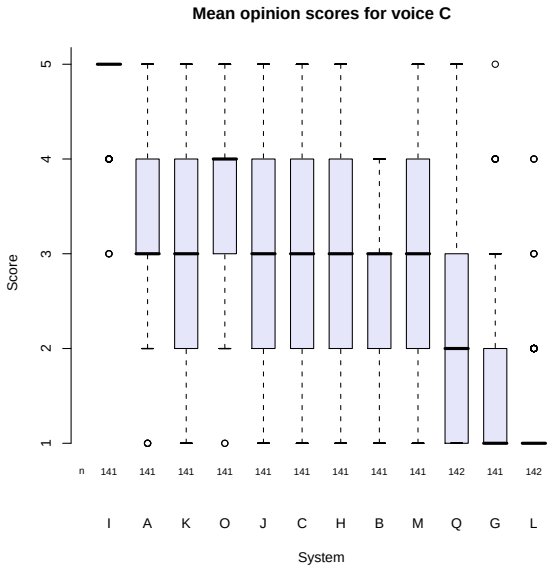

Figure 3: Boxplot showing similarity scores between systems and the original speaker for voice C, which was built from the system builder selected data set. System I represents natural speech.

Whereas it is inappropriate to compare means on such scales, it is entirely appropriate to compare medians. For this reason, this paper discusses the results in terms of medians. One outcome of this analysis is that it may be harder for the reader to distinguish between the performance of some pairs of systems by visual inspection of either descriptive statistics just involving medians, or of boxplots of the data. This is entirely intentional, because the conclusions that may be reached by inspecting **means** are statistically unfounded. For instance, the 'clear winners' that may arise when observing the highest means (opposed to those that arise through proper statistical inference) are not generally correct. We shall see that the inferential statistics qualify this presentation and that the medians do show the true picture and that there often are not statistically significant differences between systems. It is essential to take this into account when making claims about whether one system is 'better' or 'worse' than any other system. Claims based on a ranking of the means are often going to be false.

1.2. Multi-dimensional scaling analysis

One of the drawbacks of the technique of asking subjects to make judgements of a synthetic utterance on a scale of naturalness, intelligibility, similarity, etc. is that it is difficult to determine exactly what cues subjects use to make their judgements.

To investigate this issue, the evaluation included in section 2 a task where pairs of utterances from different systems were presented to subjects who were asked if the quality was the same or different. The results from this task can be compiled into a dissimilarity matrix and then analysed using multi-dimensional scaling (MDS) technique.

MDS techniques take as their input *proximity values*—that is, numbers that indicate "how similar or how different two objects are, or are perceived to be" [4]. The output of an MDS analysis is an n-dimensional *stimulus space* or map, in which each object—here, each of the synthesis systems—is represented by a single point. The key characteristic of an MDS map is that the relationship between the inter-point distances on the map, and the inter-object proximity values is such that two objects that are physically or psycho-physically close are represented by two points that are close on the map, while two objects that are physically or psycho-physically distant are represented by two points that are farther apart on the MDS map. MDS analysis determines the configuration of the points within the stimulus space by "minimis[ing] the disparity between the Euclidean distances given the dissimilarity matrix [i.e., the proximity data] and the Euclidean distances in the object space, in the least squares sense" [5][p.2169]. Additionally, an objective measure called *stress* can be used to determine how well the configuration of the points in the stimulus space represents the proximity values. The stress of any given configuration is the square root of a normalised residual sum of squares [4]; the smaller the stress, the better the fit of the configuration to the proximity data.

A second characteristic of an MDS map is that the dimensions that make up the space often correspond (directly or indirectly) to the physical or psycho-physical dimensions used most heavily by subjects to make their proximity judgements. An MDS map can have any number of dimensions: These are specified by the user before analysis is carried out. However, as the goal of most MDS analysis is to reduce the complexity of a given data set, very large numbers of dimensions are rarely specified: As noted by [4], "It is not useful to examine only a configuration with so many dimensions that you cannot comprehend it" [p.58]. Most MDS analyses, therefore, specify the smallest number of dimensions possible while achieving a low stress value. Often this is done by gradually increasing the number of dimensions just to the point at which the stress measure stops decreasing and levels off—i.e., the point at which adding further dimensions does not give a better fit of the map to the dimensions [4].

By examining and interpreting both the MDS map dimensions and the configuration of the points within those dimensions, it should be possible to determine the underlying characteristics of the objects represented in the space that led to subjects' responses to those objects. For Blizzard, this means that analysis of the MDS space should allow us to identify some of the acoustic characteristics of different synthesis systems that relate to listeners' judgements of the speech they produce. Analysis and interpretation of an MDS stimulus map can be done in a number of ways. For many two- and possibly three-dimensional spaces it may be possible to interpret the space by visually examining the distribution of the objects within the space and trying to find any underlying pattern(s) in the organisation of the points. For stimulus spaces with four or more dimensions (i.e., numbers of dimensions that are less straightforward to represent in a way that can be examined visually) or for more complex two- and three-dimensional maps, it is often necessary to make use of other methods, such as multiple regression, to regress relevant variables on the coordinates of the points in the MDS map.

1.3. Word error rates

Section 5 of the evaluation presented subjects with a Semantically Unpredictable Sentence (SUS) task where subjects were asked to type in what they heard. A word error rate score for each stimuli is calculated. This scale is an interval scale which allows us to, at least in theory, compare means.

2. Results

The results for each section of the task are now presented individually. To increase the likelihood of meaningful statistical results, the main inferential results presented are calculated

by combining categories of variables together where it is reasonable to do so. For example, previous Blizzard evaluations [6] have shown that the differences in response between different listener types are minimal, so the listener types have been combined in our main analysis. The original intention was to perform a direct comparison between U.K. and U.S. undergraduates, but the limited number of U.S undergraduate listeners has not made this possible.

The structure of the analysis was designed to provide an overall picture of the general performance of the different systems. It may be possible to show that a particular system is better than some other system in some specific situation, by only analysing a small subset of the data, but this is against the investigative (and not competitive) spirit of the Blizzard Challenge; it is also difficult to justify with the small numbers of data points which result.

In general the results are presented first for voice A (the voice using the full data set) and then for voices B and C (voices using the smaller data sets)

Standard boxplots are presented for the ordinal data where the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. Bar charts are presented for the word error rate interval data.

2.1. Sections 3 & 4: Mean opinion scores

We first discuss the combined data from sections 3 and 4 of the evaluation. As the two tasks evaluate utterances that can be considered 'in domain' for the full voice, i.e. there was a large portion of both news and conversation data available to built the voice from, the results from these two sections has been pooled to increase the value of n (i.e., the number of data points) for each group. Table 1 shows descriptive statistics for the mean opinion scores for voice A.

System	Median	MAD	mean	sd	n	NA
A	4	1.5	3.8	0.92	260	140
B	3	1.5	3.0	0.90	261	139
C	3	1.5	3.2	1.01	260	140
D	2	1.5	2.6	1.03	261	139
E	3	1.5	3.0	0.94	261	139
F	1	0.0	1.5	0.75	260	140
G	1	0.0	1.4	0.71	261	139
H	3	1.5	3.2	1.01	260	140
I	5	0.0	4.7	0.58	262	138
J	3	1.5	3.4	1.04	261	139
K	4	1.5	3.6	0.91	261	139
L	1	0.0	1.3	0.62	260	140
M	3	1.5	3.0	1.05	261	139
N	3	1.5	2.7	1.00	260	140
O	4	1.5	3.5	0.93	260	140
P	4	1.5	3.9	0.95	261	139
Q	2	1.5	2.5	0.90	260	140

Table 1: Mean opinion scores for voice A (full data set) on the combined results from sections 3 and 4 of the evaluation. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and NA (data points excluded due to missing data)

As discussed in section 1.1, the mean values here cannot be compared in a statistically meaningfully way. However, the MOS means have been used from this point forward to produce an ordering for systems shown on the plots that follow. The ordering is used to make the graphs more intuitively readable rather than to show that a particular system is the best. The ordering is **not a ranking**: System X coming before system Y in this ordering does not in any way imply that system X is significantly better than system Y.

Figure 1 displays the results of the combined MOS tests graphically. We see that natural speech (system I) has a median of 5 followed by a group of four systems (P, A K and O) with a median of 4 and group of seven systems (J, C, H, B, M, E and N) with a median of 3, a group of two systems (D and Q) with a median of 2 and a group of three systems (F, G and L) with a median of 1.

As this data is ordinal[1], to determine whether there are significant differences between the MOS scores of systems we use a series of Bonferoni-corrected pairwise Wilcoxon signed rank tests. Table 2 shows significant differences between systems with $\alpha = 0.01$. If we examine the significant differences between adjacent systems in the ordering based upon the mean MOS score for a system we only see significant differences between system I (natural speech) and system P, between systems E and N and between systems Q and F.

Figures 2 and 3 show the MOS results for voices B and C respectively. The general trend is that the MOS scores for voices B and C are lower than those for voice A, and most versions of voice C are no better than that systems voice B. Further analysis of this is beyond the scope of this paper and is left to others; more detailed information about the text selection method used in each system is probably necessary to conduct a meaningful analysis.

2.2. Section 1: Similarity to the original speaker

The results of the task asking subjects to judge how similar a system is to the original speaker are shown in Figure 4. The trend is very similar to the MOS data with no system receiving a median score matching natural speech, but there are more systems with a median score of 4. Figures 5 and 6 show the results for voices B and C respectively. Again the trend is that these voices are judged generally less similar to the original speaker than voices built on the full data set.

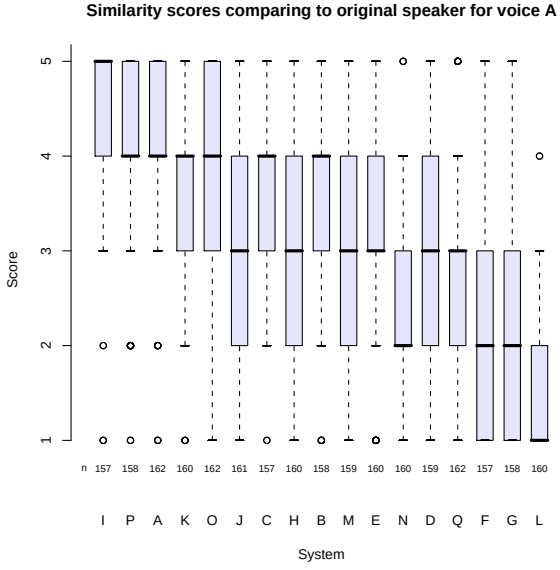

Figure 4: Boxplot showing similarity scores between systems and the original speaker for voice A, which was built from the full data set. System I represents natural speech.

[1] Even if we considered this data to me interval data, it does not meet the normality requirements to run parametric statistics

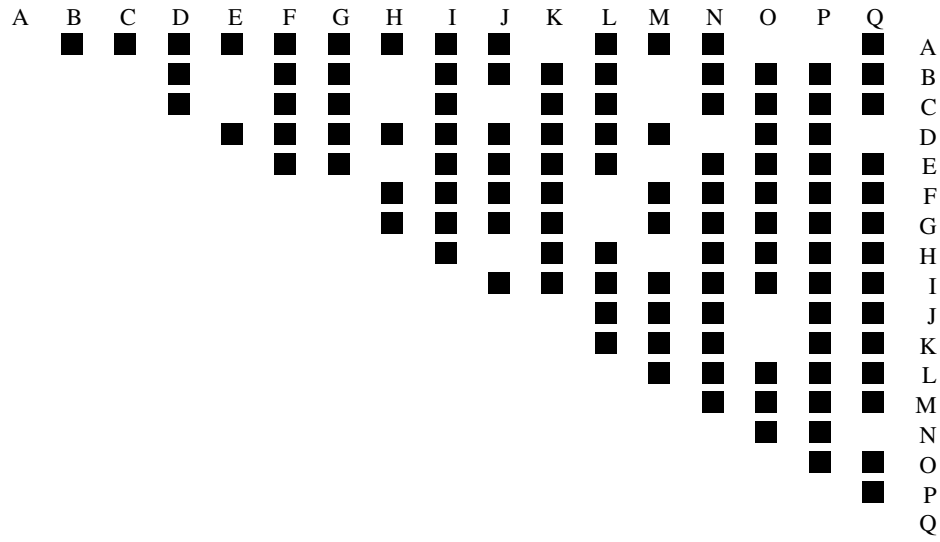

Table 2: Results of pairwise Wilcoxon signed rank tests between systems mean opinion scores, ■ shows a significant difference between systems

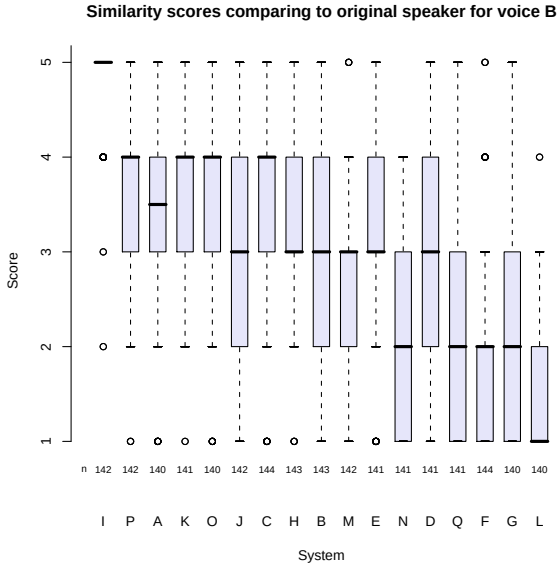

Figure 5: Boxplot showing similarity scores between systems and the original speaker for voice B, which was built from the Arctic data set. System I represents natural speech.

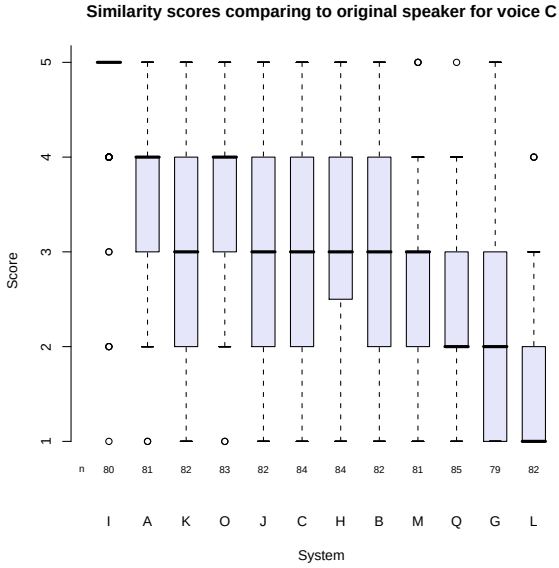

Figure 6: Boxplot showing similarity scores between systems and the original speaker for voice C, which was built from the self selected data set. System I represents natural speech.

2.3. Section 5: Word error rates

Figures 7, 9 and 9 show the word error rates for the three voices. The trends found in the MOS and similarity data continue here, with voice A generally performing better than voices B and C. If the word error rate data are analysed in terms of native and non-native listeners, there are quite large discrepancies between the word error rates that are likely to be statistically significant, although this has not been tested here. Because any target audience of a speech synthesis system is likely to be a combination of native and non-native listeners, it seems appropriate to present the combined results.

2.4. Section 2 - Multi-dimensional scaling

PROXSCAL Multidimensional scaling was performed on the dissimilarity matrix produced from the results of section 2 of the evaluation. The representation discussed here is presented in two dimensions. The two dimensional analysis results in a stress value of 0.572, increasing the number of dimensions to three only decreases the stress to 0.458. It is still possible that the third dimension will reveal additional information about what listeners are attending to, but this is left as future work.

Figure 10 shows the resulting space, which is displayed with natural speech towards the origin, but the rotation of the axes is arbitrary. Playing samples from different systems in lines parallel to the notional axes suggests that the dimensions of this data relate in some way to a global unnaturalness of utterances along the x-axis, which we will call 'roboticity'

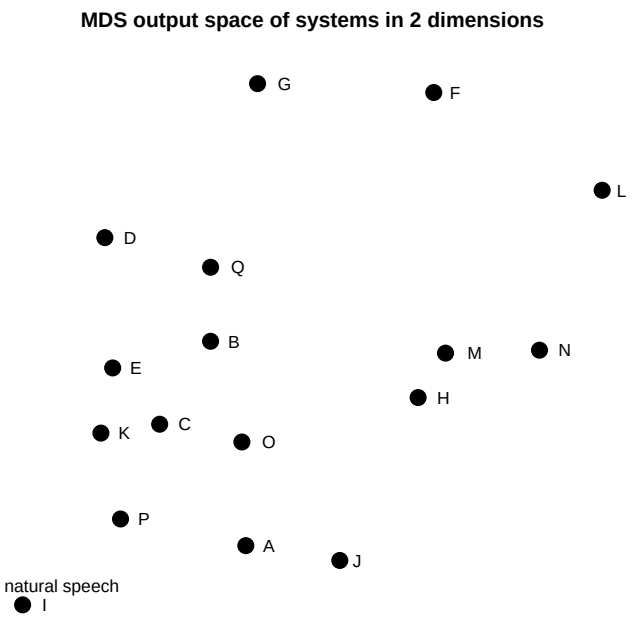

Figure 10: Two dimensional mutli-dimensional scaling output space for voice A.

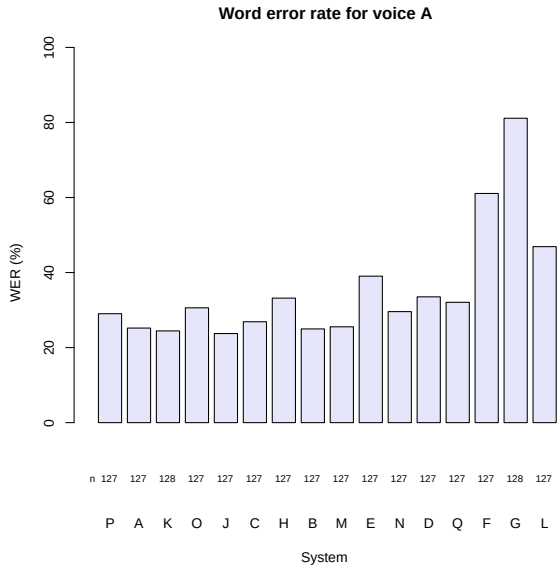

Figure 7: Bar chart showing system word error rates for voice A, which was built from the full data set.

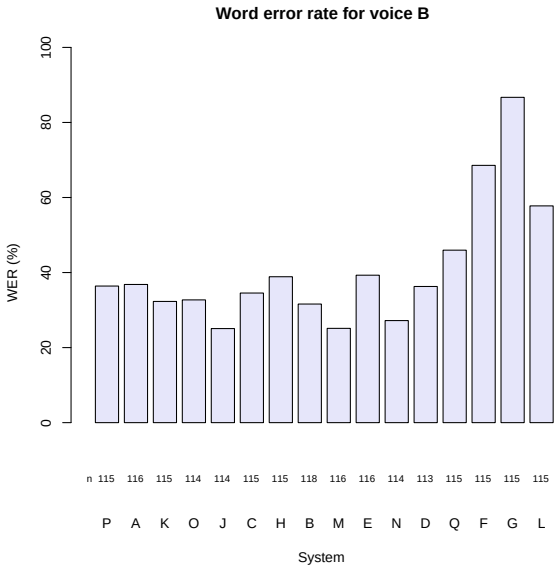

Figure 8: Bar chart showing system word error rates for voice B, which was built from the ARCTIC data set.

and local problems with naturalness which we will call 'join-discontinuity' along the y-axis, although the current axis rotation is arbitrary and could potentially be fine tuned by further analysis and experimentation.

Correlations between the coordinates with the current axis rotation in this space and the other test statistics are presented in table 3.

It is first interesting to note that there is no significant correlation between the x and y values, demonstrating that these values can be considered independent factors. Similarity judgements correlate well with both the x and y values suggesting that subjects consider both of these potential factors when judging similarity to natural speech.

Word error rate only correlates with the y value 'join discontinuity'. Suggesting that a voice does not necessarily have to sound natural (at least in the 'robotic' sense) to be intelligible.

Mean opinion score again correlates strongly with the

	x	y	Sim	MOS	WER
x			$r = -.775$ $p < 0.001$		
y			$r = -.766$ $p < 0.001$	$r = -.921$ $p < 0.001$	$r = .846$ $p < 0.001$
Sim	$r = -.775$ $p < 0.001$	$r = -.766$ $p < 0.001$		$r = .864$ $p < 0.001$	$r = -.724$ $p < 0.001$
MOS		$r = -.921$ $p < 0.001$	$r = .864$ $p < 0.001$		$r = -.824$ $p < 0.001$
WER		$r = .846$ $p < 0.001$	$r = -.724$ $p < 0.001$	$r = -.824$ $p < 0.001$	

Table 3: Significant Pearson correlations between the MSD space coordinates and mean opinion scores, similarity judgements and word error rates.

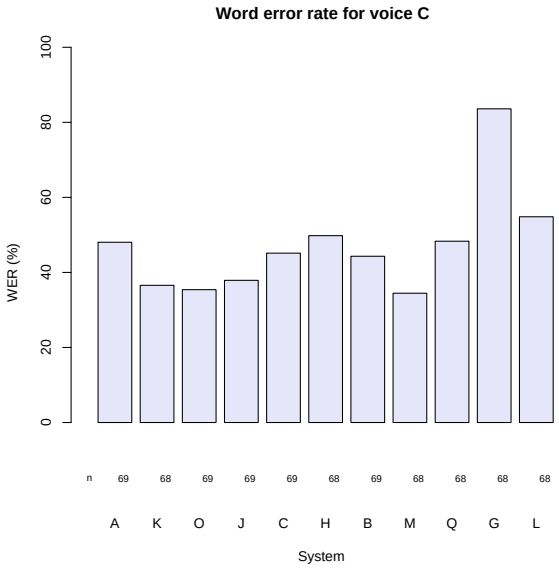

Figure 9: Bar chart showing system word error rates for voice C, which was built from the self selected data set.

y value ('join-discontinuity'), as would be expected. There is however only a weak correlation (not shown above, $r = -.513, p < 0.05$) with the x value ('roboticity').

To further confirm our interpretation of these results we are currently conducting an experiment to discover if naive subjects can make judgements on a Likert-type scale to specifically rate the factors of 'roboticity' and 'join-discontinuity' (if suitably described) and to see if we can achieve stronger correlations with the above axes than tasks performed so far.

3. Discussion

Each section of the evaluation has produced a useful outcome. In general, subjects used the full ranges of the scales they were asked to use. This may have been helped by the ordering of the tasks which allowed them to become familiar with both natural speech references and synthetic examples from all systems, at the beginning of the test (in section 1).

The Multi-dimensional scaling produced interesting results that reflected the other results although there is still further analysis that can be performed here. For example, the orientation of the resulting MSD space is arbitrary and an attempt could be made to find the optimal rotation where the axes maximally correlate with other factors, using a technique such as principal component analysis.

There is also scope for performing the MDS evaluation in higher dimensional spaces: [7] suggests that prosodic cues may affect listener judgements of synthetic speech, but this is not evidenced in our two dimensional MDS space.

4. References

[1] Mark Fraser and Simon King, "The blizzard challenge 2007," in *Proc. Blizzard Workshop (in Proc. SSW6)*, 2007.

[2] Rensis Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, pp. 1–55, 1932.

[3] M Marcus-Roberts and F. S Roberts, "Meaningless statistics," *Journal of Educational Statistics*, vol. 12, no. 4, pp. 383–394, 1987.

[4] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, Sage University Paper series on Quantitative Applications in the Social Sciences. Sage Pubns., Beverly Hills and London, 1978.

[5] J. L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *Journal of the Acoustical Society of America*, vol. 110, pp. 2167–2182, 2001.

[6] Alan W. Black and Keiichi Tokuda, "The Blizzard Callenge - 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc Interspeech 2005*, Lisbon, 2006.

[7] C. Mayo, R. A. J. Clark, and S. King, "Multidimensional scaling of listener responses to synthetic speech," in *Proc. Interspeech 2005*, Lisbon, Portugal, Sept. 2005.