

# The Blizzard Challenge 2006

*Christina L. Bennett and Alan W Black*

Language Technologies Institute  
Carnegie Mellon University, Pittsburgh, PA, USA  
{cbennett, awb}@cs.cmu.edu

## Abstract

Last year the Blizzard Challenge 2005 introduced the speech synthesis community to the concept of large scale, multi-site evaluation of TTS systems using common data. In this, the second year of the Blizzard Challenge, we again tackled this task. Participation increased dramatically, out of a total of 17 initial sites that showed interest, a total of 14 sites from around the world actually submitted entries. In this paper we discuss the results, difficulties, and differences in this year's Challenge.

**Index Terms:** speech synthesis, evaluation, corpus based synthesis

## 1. Introduction

Everyone understands the importance of evaluation. What is often not understood is the importance of many varied types of evaluation. Most groups conduct some sort of in-house evaluation periodically to determine whether the changes being made to their systems are actually improvements. In the field of speech synthesis, this had often been the extent of it. With each group using different datasets, it was very hard to evaluate how well particular techniques worked, as it is clear that the database itself is an important contributor to the overall quality of a corpus-based synthetic voice. Last year six sites took part [1]. The challenge clearly focused groups, and last year's somewhat unexpected results highlighted that evaluation on common datasets is a critical aid in our continued goal of better quality speech output.

The Blizzard Challenge is to take a provided single-speaker database of recorded speech plus its transcription and build a synthetic voice from it. Five sets of 50 sentences must then be synthesized with the constructed voice. These synthesized utterances are then listened to by three different listener groups.

The data for Blizzard Challenge 2006 was provided by ATR, and consists of some 5000 phonetically balanced utterances spoken by a male speaker of US English. Text, automatically derived phone labels, and Festival Utterance Structures were additionally provided. The data includes CMU ARCTIC [2], news stories, and conversational text designed for machine translation systems (BTEC [3]). Following desires expressed in the previous year, this database is substantially bigger than that made available in 2005, which consisted of only CMU ARCTIC, though four different voices were provided.

Also following last year, the five genres were kept the same. These were novels, news, conversation, phonetically confusable sentences, and semantically unpredictable sentences (SUS). The novels, news, and conversations test sentences were held out from the full ATR dataset before we released the database to the participants, specifically this allowed us to have natural speech

examples for those test sentences, which provides both a benchmark for our synthetic examples and a method to detect listeners who are not treating the task seriously

This marks the second year of the Blizzard Challenge. The success of the Blizzard Challenge 2005 led to markedly increased participation this year. More than twice as many sites from both industry and academia participated, representing multiple labs from Asia (7), the United States (2), and Europe (5).

- ATR, Japan
- Cereproc, Edinburgh, UK
- Carnegie Mellon University, USA
- CSTR, University of Edinburgh, UK
- DFKI, Saarbrücken, Germany
- IBM, Yorktown Heights, USA
- IBM, Haifa, Israel
- iFLYTEK Research, China
- IVO, Poland
- Kyoto University, Japan
- Microsoft Research Asia, China
- Nagoya Institute of Technology, Japan
- University College Dublin, Ireland
- University of Science and Technology of China

An annual challenge of this sort is both a valuable resource and a strong motivator for healthy competition among researchers, which in turn, leads to innovation. The Blizzard Challenge 2006 was hosted by Carnegie Mellon University and conducted from late June to mid August 2006.

## 2. Blizzard evaluation methods

Much of our basic evaluation process has been carried over from last year's challenge. Refer to Bennett 2005 [4] for further details, as here we focus primarily on the differences in this year's challenge.

### 2.1. Test ordering

Similarly to last year, each listener was assigned to a group upon registration. Instead of ten groups, this year we determined the number of groups based on the number of participating sites. These group assignments were used to determine the order of systems a user would hear through each of the sentences and each of the tests. Additionally, the group assignment determined whether a listener would hear samples synthesized from the full speech dataset or the smaller Arctic-based subset. We felt it would be too labor-intensive for each listener to evaluate samples from both versions of every system for every test, thus listeners only heard samples from the full set

or Arctic set. Therefore, because there were 14 participating systems, plus the natural “system”, and two distinct datasets, there would be 30 distinct ordering groups. This is 15 groups per dataset. (Note that one of these groups was discarded in advance because one site did not submit an Arctic-based system. As a result, there were 29 ordering groups overall.)

Because of the large number of participants this year, we opted to have each listener hear only one sample per system per test. With no natural samples for the final two tests, this means that each listener would hear 14 to 15 samples per test in the first three tests and 13 to 14 samples in each of the final two tests, depending on whether they were assigned to one of the full set groups or one of the Arctic set groups.

In order to guarantee every possible ordering of systems, a Latin square was utilized. For the first three tests, a 15 x 15 Latin square was devised for samples synthesized using the full speech dataset. A 14 x 14 square was used for the first three tests of the Arctic-based samples as well as the final two tests of the full set samples. Similarly, a 13 x 13 square was used for the final two tests of the Arctic-based samples. Using a Latin square design means that each system had its sample in each position (first, second, third, etc.) exactly once, *i.e.* in a single group. Sentence order was maintained across groups but system order varied such that each system held each position. Additionally, we attempted to use a balanced Latin square, meaning that surrounding context is also maximally varied (*i.e.* A follows B only once and vice versa); however, it should be noted that balanced Latin squares do not exist for odd-numbered squares [5]. As a result, our Latin squares were nearly, though not completely balanced.

Based on the number of ordering groups established above, we hoped to attain at least five listeners per group. With 29 groups and three different types of listeners (described in section 2.2 below), this meant our target number of listeners was 145 per listener group and 435 listeners overall. Initially listeners were assigned to groups in a rolling fashion, adding one listener to each group in order of registration and based on listener type; however, toward the end of the evaluation this assignment method was modified such that the groups with the fewest number of listeners who had actually completed all of the tests were assigned new listeners first. This was an attempt to balance the number of listeners with completed tests among all of the groups, thus balancing the collected data for all of the systems.

## 2.2. Listener groups

Once again we used three categories of listeners in the evaluation – speech synthesis experts (Group S), volunteers (Group V), and native U.S. English speakers under the age of 25, loosely termed “undergraduates” (Group U).

Participating sites were asked to provide 10 speech synthesis experts as listeners. These made up group S and were predictably from around the world with varied linguistic backgrounds. This group has a high level of motivation to complete the task because of their professional ties and interest. Over the course of the evaluation, 134 type S listeners registered. This is 11 short of our desired number of listeners, which we were hoping to achieve despite the fact that had every site provided their 10 listeners, we still would have been short by five. In any case, 126 of the 134 registered completed some

portion of the tests, while 83 completed all of the tests. For this group, the large number of non-native listeners had an impact on their ability to complete the open response tests, bringing down the completion number significantly.

Group V was composed strictly of volunteers who had heard about the challenge from a message board or list email post, by word of mouth, or by other such means. Members of this group have very little to gain (or lose) from participating, thus they displayed the least motivation to complete all of the tests. An encouraging 214 listeners were registered under group V. While 174 completed some of the tests, only 113 completed all of them. Thus for this group we were only 32 listeners short of reaching the overall goal of 145 completed. It is worth noting though, that there is some evidence to suggest that some of these listeners should have been included in the S group and may have registered as type V in error.

Group U was the most demographically controlled of the three groups but also the most difficult group to populate. This group was restricted to native U.S. English speakers, as this is the dialect of the voice talent and thus the voices being tested. Members of this group were typically undergraduate students, though any native speakers under the age of 25 were eligible. In order to help populate this group, U listeners were paid \$10 in the form of an Amazon gift certificate. Last year we found that a \$5 payment was insufficient to draw significant numbers. This year, several other factors contributed to the difficulty of gathering a sizeable number of listeners in group U. These issues will be discussed in section 4. We were only able to amass 55 registered listeners under group U, well short of the goal outlined in the previous section. Fifty completed some portion of the tests, but only 44 completed all of the tests.

## 2.3. Test types

Test types were identical to that of the Blizzard Challenge 2005. Tests 1 through 3 were Mean Opinion Score (MOS) tests [6] with a scale of 1 to 5, where a score of 5 is best. The final two tests were Modified Rhyme Test (MRT) [7] and Semantically Unpredictable Sentences (SUS) [8]. These are open response tests where the listener is asked to type the words that they hear into a text box, rather than provide a score. For these open response tests, word error rate (WER) is calculated from the listeners’ inputs.

## 2.4. Test design

For this year’s Challenge, we again chose to conduct an entirely online evaluation. The number and variety of listeners available are dramatically increased by conducting a study online, though the experimenters have considerably less control over the environment and likelihood of completion of the evaluation. The same evaluation software, developed last year, was modified to suit this year’s conditions including having only one speaker from which to synthesize samples, testing two distinct speech datasets (the full set and a subset), and including a vastly increased number of participant systems.

As stated previously, the evaluation was again composed of five tests, from the genres news, novels, conversation, MRT, and SUS. Each test contained 13 to 15 samples as detailed in section 2.1. All other aspects of testing were identical to last year; see [4] for details.

### 3. Results and discussion

For the purpose of anonymity, participating sites have been assigned letter identifiers, A through N. The letter O is used to denote the natural speech reference examples. Including “system O” allows us to compare natural speech recorded from the voice talent directly to the synthetic speech of the participating systems. Since reference samples were not available for MRT and SUS sentences, we are unable to compare WER scores and thus will report natural MOS scores only.

There are many dimensions along which to compare this year’s results. There are the results when using the full speech dataset versus when using the Arctic subset. There are results when the data is restricted in the strictest sense versus a more lax restriction (explained below in 3.1). And of course, there are the dimensions of listener type and test type. Comparisons could also be made to overall results of each test from last year, particularly for the Arctic subset. Age, gender, native tongue, and more are also available for scrutiny. Here we provide highlights but will by no means cover all of these potential comparison points.

#### 3.1. Data restrictions

When conducting an evaluation of this size, often there are some data that must be excluded. The following were identified as reasons for exclusion:

- incomplete tests,
- failure to follow directions,
- inability to respond to type-in tests (*i.e.* language barrier),
- unusable responses, such as those lacking effort, inappropriate to the task, or extremely contrary to expectation.

These conditions are described further in [4]. The primary differences in the two data restrictions were as follows. In the lax case, any incomplete (*i.e.* unfinished) test was discarded, though other complete tests from the same listener were preserved. In the strict case, if a listener did not complete all of the tests in their entirety, all data from the listener was discarded. This was in an effort to ensure the same number of inputs for each of the tests. Also, in the strict case, if any test contained three or more “missing” entries, the listener was discarded. Here we use “missing” entries to refer to the case where, often as a result of poor English skill, a listener may have given responses to most type-in samples, though not all. If they could not discern any words for three or more of the samples in a single test, they were removed under this criterion. In the lax case these listeners were left in so long as they gave some legitimate responses because they showed effort in completing the task.

#### 3.2. Strict results

These results were calculated using the strictest data restrictions. These are considered to be the official results and were distributed to the participating sites.

In Table 1 we see the results of samples built using the full speech dataset. These results are divided by listener type and ordered by system performance. Average MOS over the three MOS tests and WER for the two open response tests are given. Remember that for the MOS tests, a 1 to 5 scale was used, thus

the closer the score is to 5, the better. For WER, a lower score is better because it means listeners were better able to discern the words being spoken by the system. The score given represents the number of errors found in listener inputs for each of the systems.

S		V		U	
MOS	type-in	MOS	type-in	MOS	type-in
O - 4.659	n/a	O - 4.514	n/a	O - 4.441	n/a
K - 3.696	C - 14.63	C - 3.514	C - 20.48	K - 3.738	M - 11.90
H - 3.400	I - 17.22	K - 3.458	A - 23.59	C - 3.726	K - 12.50
L - 3.370	A - 17.41	M - 3.220	I - 24.72	H - 3.536	C - 14.58
C - 3.319	M - 18.33	L - 3.203	K - 25.14	M - 3.441	H - 15.18
M - 3.252	G - 18.70	H - 3.170	M - 25.42	L - 3.381	I - 15.48
G - 3.163	L - 20.00	G - 3.124	H - 26.41	I - 3.369	A - 17.56
I - 3.089	H & K - 21.11	I - 3.017	G - 26.69	G - 3.238	G - 19.05
D & F - 2.948	D - 23.15	A - 3.000	L - 27.12	A - 2.941	B & L - 19.35
A - 2.926	N - 23.89	F - 2.633	B - 31.92	F - 2.738	D - 19.64
N - 2.519	B - 24.07	B - 2.458	N - 33.90	B - 2.560	N - 24.70
B - 2.467	J - 27.96	N - 2.441	J - 35.31	N - 2.536	J - 32.14
J - 2.000	F - 32.41	J - 1.944	F - 43.64	J - 2.012	F - 37.80
E - 1.393	E - 49.44	E - 1.571	E - 61.16	E - 1.619	E - 53.27

Table 1. *FULL-STRICT: Systems ranked by performance on the full dataset with strictest data restrictions.*

S		V		U	
MOS	type-in	MOS	type-in	MOS	type-in
O - 4.675	n/a	O - 4.617	n/a	O - 4.792	n/a
C - 3.246	C - 18.64	C - 3.636	C - 18.52	C - 3.792	C - 10.94
K - 3.070	L - 20.39	L - 3.469	L - 22.22	K & L - 3.271	L - 19.79
L - 3.035	I & M - 21.27	K - 3.253	I - 23.30	H - 20.31	
M - 3.009		H - 3.124	M - 23.92	M - 3.208	A - 20.83
H - 2.719	H - 23.46	M - 3.074	H - 24.38	H - 3.167	M - 21.35
I - 2.667	A - 25.22	I - 2.969	A - 25.15	D - 3.021	I - 24.48
A - 2.588	G - 29.17	G - 2.821	K - 29.94	A - 2.958	K - 29.17
D - 2.579	K - 31.80	A - 2.753	G - 30.71	I - 2.854	G - 30.73
F - 2.430	J - 33.11	D - 2.741	D - 32.72	G - 2.750	D - 32.29
G - 2.368	D - 34.65	F - 2.327	J - 37.35	F - 2.417	B - 33.85
B - 1.921	B - 34.87	B - 2.284	B - 39.35	B - 2.167	J - 41.67
J - 1.842	F - 37.06	J - 1.994	F - 41.98	J - 1.833	F - 42.19
E - 1.316	E - 52.41	E - 1.426	E - 50.31	E - 1.500	E - 54.69

Table 2. *ARCTIC-STRICT: Systems ranked by performance on the arctic subset with strictest data restrictions.*

A discussion of the results here as well as the lax results presented in 3.3 will be given in section 3.4 below.

### 3.3. Lax results

Results in this section were calculated using more lax data restrictions. These results are included for comparison because of the data sparseness of the strict set. Keep in mind they may contain more non-native listeners, and the number of subjects per test are not equal. Generally speaking, the number of subjects decreases in order of the tests (*i.e.* Test 1 has more than Test 2, etc., up to Test 5, the SUS test).

S		V		U	
MOS	type-in	MOS	type-in	MOS	type-in
O - 4.583	n/a	O - 4.508	n/a	O - 4.490	n/a
K - 3.632	C - 19.09	K - 3.521	C - 21.77	K - 3.739	M - 11.73
L - 3.456	A - 21.36	C - 3.496	A - 24.51	C - 3.685	K - 12.61
C - 3.353	I - 21.52	L - 3.252	I - 26.21	H - 3.511	C - 14.66
H - 3.328	M - 21.84	H - 3.223	K - 26.60	M - 3.424	H & I - 15.25
M - 3.230	G - 22.49	M - 3.207	M - 26.86	L - 3.391	
G - 3.211	L - 23.30	G - 3.128	G - 27.38	I - 3.380	A - 17.60
I - 3.098	H - 23.95	I - 3.074	H - 27.64	G - 3.239	G - 18.77
F - 3.049	K - 24.60	A - 3.021	L - 27.90	A - 2.957	B & L - 19.06
A - 2.995	B & D - 26.86	D - 2.822	D - 32.20	D - 2.826	
D - 2.946		F - 2.690	B - 32.46	F - 2.696	D - 19.35
N - 2.495	N - 27.02	B - 2.488	N - 34.68	B - 2.554	N - 24.63
B - 2.456	J - 31.23	N - 2.422	J - 36.11	N - 2.467	J - 32.26
J - 1.961	F - 34.63	J - 1.934	F - 44.20	J - 2.011	F - 37.24
E - 1.431	E - 51.29	E - 1.550	E - 61.54	E - 1.609	E - 52.49

Table 3. FULL-LAX: Systems ranked by performance on the full dataset with lax data restrictions.

S		V		U	
MOS	type-in	MOS	type-in	MOS	type-in
O - 4.663	n/a	O - 4.568	n/a	O - 4.800	n/a
C - 3.338	C - 22.37	C - 3.472	C - 20.17	C - 3.909	C - 12.50
K - 3.263	L - 24.51	L - 3.336	L - 23.37	L - 3.346	L - 21.30
L - 3.219	I - 24.90	K - 3.214	M - 25.45	K - 3.309	M - 21.76
M - 3.013	M - 25.49	H - 3.039	I - 25.87	H & M - 3.273	A & H - 22.22
H - 2.863	H - 26.65	M - 2.948	H - 26.84		
I - 2.738	A - 29.18	I - 2.843	A - 27.12	A - 3.109	I - 26.85
A - 2.700	K - 33.27	G - 2.734	K - 31.15	D - 3.055	G - 30.56
D - 2.663	G - 33.66	A - 2.712	G - 31.43	I - 2.873	K - 31.94
F - 2.656	B - 36.58	D - 2.651	D - 34.91	G - 2.855	B - 33.80
G - 2.550	J - 36.77	F - 2.445	J - 38.53	F - 2.418	D - 34.26
B - 1.969	D - 37.74	B - 2.170	B - 40.19	B - 2.218	J - 43.52
J - 1.888	F - 38.72	J - 1.917	F - 42.84	J - 1.927	F - 43.98
E - 1.381	E - 52.53	E - 1.397	E - 52.99	E - 1.636	E - 54.63

Table 4. ARCTIC-LAX: Systems ranked by performance on the Arctic dataset with lax data restrictions.

As in the previous section, Table 3 on the left shows MOS and WER results for each of the listener types. Results are ordered by system performance on the full speech dataset. Table 4 then shows the same for systems created using only the Arctic subset of the speech data.

### 3.4. Discussion of results

In this section we provide a general discussion of the results presented. The additional tables, 5 and 6, are included for ease of comparison across systems. Table 5 presents the same results as in Tables 1 and 2, *i.e.* strict results from both the full and the Arctic datasets. Table 6 then presents the results found in Tables 3 and 4, which were lax results from both dataset configurations. Unlike their counterparts, results in these tables are sorted *by system* for easier analysis of each system across datasets, test types, and listener types.

Generally speaking, most systems demonstrated better performance when using the full speech dataset than when using the significantly smaller Arctic subset of that data. In other words, WER was lower and average MOS was higher for the samples built from the full set compared to those built from the Arctic subset. Whether the strict or the lax data restrictions were used, this was true for all systems as rated by the S listeners; the other listener groups were slightly less consistent in this regard. However, for one team, nearly the opposite case was true. Team C emerged as a clear winner in both WER and average MOS when Arctic-based systems were compared, often performing as well as or better than their own system trained on the full speech dataset.

Contrary to the trend among the majority of synthetic speech systems, when the natural speech samples as rated in the Arctic MOS tests versus the full set MOS tests are compared, we see a very different result. Keep in mind that the natural samples, because they are natural, are *identical* under both testing conditions. In this case, we see a lower average MOS when tested with full set systems versus when tested with Arctic subset systems. In some cases the difference was very slight, but the trend was observed across the board – for all listener types and both levels of data restriction. One possible explanation involves the fact that most systems performed better on the full set. Keep in mind that every scale is relative, and listeners only scored Arctic systems *or* full systems. In the case of the full-based tests, overall quality was generally quite good. In the case of the Arctic-based tests, quality was somewhat reduced. Thus, by comparison, the natural samples sounded consistently that much better than their synthetic counterparts. Unfortunately we do not have WER information for the natural samples, which would be very interesting to compare given this observation. Also, perhaps Team C’s slightly opposite-of-the-trend results (often performing better on the Arctic set than on the full set) can also be attributed to this phenomenon? On the other hand, it’s possible that this team has simply been more successful at making full use of sparse data.

It is less clear which system did best using the full speech data. Team C again did well with the best WERs under lax restrictions and other top ranks; however, Team K had just as many top finishes on the full set. (Team M also grabbed the top spot under two of the conditions.)

Last year we observed that type S listeners were not only better at understanding synthetic speech (lower WERs) but that

they also liked it more than the other populations. This year we noticed a change. S listeners often performed better on the open response tests than their type V counterparts, but on the whole, the type U listeners had higher MOS averages and better WERs, particularly for the best systems, reaching an impressive 10.94 WER on the best performer for the strict Arctic set and an 11.9 WER on the full set. We'll discuss how this population is somewhat different from last year in Section 4.2.

As was the case last year, on average female listeners outperformed male listeners on the open response task. Average MOS was also slightly higher for the female population. Of the MOS tests, the news genre was again the lowest performer. Standard deviations are slightly larger this year, likely because of the division of listeners among the two test sets (full and Arctic) which meant fewer data points for each test.

S	full set		arctic subset		V	full set		arctic subset		U	full set		arctic subset	
	MOS	type-in	MOS	type-in		MOS	type-in	MOS	type-in		MOS	type-in	MOS	type-in
A	2.926	17.41	2.588	25.22	A	3.000	23.59	2.753	25.15	A	2.941	17.56	2.958	20.83
B	2.467	24.07	1.921	34.87	B	2.458	31.92	2.284	39.35	B	2.560	19.35	2.167	33.85
C	3.319	<b>14.63</b>	<b>3.246</b>	<b>18.64</b>	C	<b>3.514</b>	<b>20.48</b>	<b>3.636</b>	<b>18.52</b>	C	3.726	14.58	<b>3.792</b>	<b>10.94</b>
D	2.948	23.15	2.579	34.65	D	2.802	30.65	2.741	32.72	D	2.881	19.64	3.021	32.29
E	1.393	49.44	1.316	52.41	E	1.571	61.16	1.426	50.31	E	1.619	53.27	1.500	54.69
F	2.948	32.41	2.430	37.06	F	2.633	43.64	2.327	41.98	F	2.738	37.80	2.417	42.19
G	3.163	18.70	2.368	29.17	G	3.124	26.69	2.821	30.71	G	3.238	19.05	2.750	30.73
H	3.400	21.11	2.719	23.46	H	3.170	26.41	3.124	24.38	H	3.536	15.18	3.167	20.31
I	3.089	17.22	2.667	21.27	I	3.017	24.72	2.969	23.30	I	3.369	15.48	2.854	24.48
J	2.000	27.96	1.842	33.11	J	1.944	35.31	1.994	37.35	J	2.012	32.14	1.833	41.67
K	<b>3.696</b>	21.11	3.070	31.80	K	3.458	25.14	3.253	29.94	K	<b>3.738</b>	12.50	3.271	29.17
L	3.370	20.00	3.035	20.39	L	3.203	27.12	3.469	22.22	L	3.381	19.35	3.271	19.79
M	3.252	18.33	3.009	21.27	M	3.220	25.42	3.074	23.92	M	3.441	<b>11.90</b>	3.208	21.35
N	2.519	23.89	n/a	n/a	N	2.441	33.90	n/a	n/a	N	2.536	24.70	n/a	n/a
O	4.659	n/a	4.675	n/a	O	4.514	n/a	4.617	n/a	O	4.441	n/a	4.792	n/a

Table 5. Overall results given strictest data restrictions. MOS and WER scores for all systems on both datasets, for all user types. Best performers in each category are marked in bold.

S	full set		arctic subset		V	full set		arctic subset		U	full set		arctic subset	
	MOS	type-in	MOS	type-in		MOS	type-in	MOS	type-in		MOS	type-in	MOS	type-in
A	2.995	21.36	2.700	29.18	A	3.021	24.51	2.712	27.12	A	2.957	17.60	3.109	22.22
B	2.456	26.86	1.969	36.58	B	2.488	32.46	2.170	40.19	B	2.554	19.06	2.218	33.80
C	3.353	<b>19.09</b>	<b>3.338</b>	<b>22.37</b>	C	3.496	<b>21.77</b>	<b>3.472</b>	<b>20.17</b>	C	3.685	14.66	<b>3.909</b>	<b>12.50</b>
D	2.946	26.86	2.663	37.74	D	2.822	32.20	2.651	34.91	D	2.826	19.35	3.055	34.26
E	1.431	51.29	1.381	52.53	E	1.550	61.54	1.397	52.99	E	1.609	52.49	1.636	54.63
F	3.049	34.63	2.656	38.72	F	2.690	44.20	2.445	42.84	F	2.696	37.24	2.418	43.98
G	3.211	22.49	2.550	33.66	G	3.128	27.38	2.734	31.43	G	3.239	18.77	2.855	30.56
H	3.328	23.95	2.863	26.65	H	3.223	27.64	3.039	26.84	H	3.511	15.25	3.273	22.22
I	3.098	21.52	2.738	24.90	I	3.074	26.21	2.843	25.87	I	3.380	15.25	2.873	26.85
J	1.961	31.23	1.888	36.77	J	1.934	36.11	1.917	38.53	J	2.011	32.26	1.927	43.52
K	<b>3.632</b>	24.60	3.263	33.27	K	<b>3.521</b>	26.60	3.214	31.15	K	<b>3.739</b>	12.61	3.309	31.94
L	3.456	23.30	3.219	24.51	L	3.252	27.90	3.336	23.37	L	3.391	19.06	3.346	21.30
M	3.230	21.84	3.013	25.49	M	3.207	26.86	2.948	25.45	M	3.424	<b>11.73</b>	3.273	21.76
N	2.495	27.02	n/a	n/a	N	2.422	34.68	n/a	n/a	N	2.467	24.63	n/a	n/a
O	4.583	n/a	4.663	n/a	O	4.508	n/a	4.568	n/a	O	4.490	n/a	4.800	n/a

Table 6. Overall results given lax data restrictions. MOS and WER scores for all systems on both datasets, for all user types. Best performers in each category are marked in bold.

## 4. Lessons for the future

### 4.1. Incentive

For every endeavor requiring human participation there is the question of incentive. Type S listeners have clear incentives. They get to participate in an activity to better their field and at the same time fulfill their own curiosities about the achievements of their competitors and colleagues. To a lesser degree, this may also hold true for type V listeners, who are perhaps friends, family, colleagues, or someone who in some feels a connection to the goal of improved speech synthesis. Curiosity and helpfulness are their motivating incentives. So what of the target population, listener type U, who are most removed from problem? For the past two years, our answer to this question has been payment. But is this enough?

In the recently published best-seller, *Freakonomics* [9], the authors describe a 1970's study that inspires a similar question. It was found that blood bank donations actually *decreased* when a small payment was introduced. Suddenly, donating blood was about payment rather than contributing to the greater good. Last year we noted that a \$5 payment was not sufficient for bringing in significant numbers of type U listeners. Despite doubling the payment incentive this year, the problem actually *grew*. A few possible reasons for this phenomenon are discussed in the next section. We advise that in future incarnations of the Blizzard Challenge, these issues be carefully considered and addressed before committing to a strict timeline.

### 4.2. Soliciting listeners

As described in the previous section, it was very difficult to fill our own quota of type U listeners this year. Last year, the Blizzard Challenge was conducted in the spring, during the traditional U.S. school year. Having it at this time allowed us access to captive groups of potential subjects. As a result, the listeners who made up last year's group U were primarily from two classes, one at Stanford and one at Ohio State. The instructors of these two classes asked their students to do the study. As a result, last year's group of U listeners was somewhat homogenous. This year's Challenge took place during the summer months when most U.S. undergraduates are away from their home universities. Thus, we sought listeners from very different sources. An online experiment site at Carnegie Mellon University was used to help bolster the number of U listeners; however, since anyone can access the site, we received several false U's (*i.e.* not fitting the specified demographic), which had to be moved to type V as they came in. Several U listeners this year came from a pre-college course for high school students being held on the CMU campus. Several others had to be solicited directly based on their demographic appropriateness for this group.

Interestingly, it appears as though a summer school class at the University of Michigan was asked to complete the evaluation. Many of these listeners, who registered as type V, could have filled roles as type U; however, we were unaware of this potential U group until after the evaluation was complete and listener questionnaire were scrutinized. This case might suggest that we should do some type of listener sorting rather than rely on the listeners to register using the appropriate URL.

However, determining who should receive payment complicates the task. Sorting listeners also makes balancing each of the ordering groups (determining which speech samples each listener is given) more challenging.

To a lesser degree, holding the evaluation during the summer also impacted our ability to fill the other listener groups as well. Since type S listeners are exclusively experts from the participating sites by definition, we had to rely on the participants to come up with enough listeners. For type R, we observed that general interest (from outside the synthesis community) had diminished this year. Many people who were willing listeners last year did not participate this year despite having knowledge of the new evaluation. In one case, a post to one popular bulletin board garnered a large influx of V listeners last year, but this year, that source was far less lucrative. These issues forced us to search out new sources of potential listeners.

### 4.3. Test design issues

One challenge we faced last year was the presence of homophones in the open response tests. Some had been caught and excluded beforehand, but dialect differences in particular were overlooked. This year we made a concerted effort to exclude any possible homophones during the selection of sentences for testing. Of course, we must still deal with typos, misspellings, and alternate spellings in each the open response tests. This year the data were pre-processed based on a small list of possible alternative spellings (*e.g.* the American "arbor" vs. the British "arbour"). We also included a spell-check module but have chosen not to include those results for comparison here.

As was the case last year, listeners were asked to submit an exit questionnaire upon completion of all five tests. The questions were essentially the same as last year – age, gender, education, and experience questions, as well as questions about the types of tests used in the study. This year we added two questions about the listener's test-taking environment, specifically, which browser they used and what type of audio output device. These were added because of certain compatibility issues people commented on last year. Comments and suggestions for improvement were also solicited. Listeners were instructed that all exit questions were optional.

Again, the overwhelming majority of comments this year were positive; however, media player issues were again the primary complaint. We suggest that for future incarnations of the Blizzard Challenge, an embedded media player be used for playing the audio files. Too often the browser preferences of the listener would cause the file to be played on a new page, advancing the browser window, rather than launching an external window. An additional advantage to integrating the media player into the page is the potential to control, or at least monitor, the number of times a listener plays each file.

Also noted this year were a few requests to keep a status tracker *throughout* the tests. In the current design, the listener's status on each of the tests is only presented on the main test page. This would be a simple modification.

Another noteworthy comment from a few listeners this year was a concern that the organizers had a hidden agenda. These listeners commented that they felt *they* were being tested rather than the systems as stated repeatedly in the evaluation description and instructions. Primarily because of the difficult

words in the SUS test combined with the exit questions about education, these listeners felt their intelligence was what was actually being studied. This is a significant problem without a clear remedy.

We have already noted the benefits of conducting a large evaluation online; however, there are definite obstacles in doing so. Several of these obstacles were more apparent this year. Firstly, when there is no experimenter present during testing, it is obviously impossible to assist the listeners when they encounter difficulties or have questions. It is also extremely challenging to ensure that listeners are appropriately assigned to each of the different listener groups, and within those groups, to ensure that the data is balanced across all testing conditions. Perhaps the largest challenge though is the difficulty in getting enough of your listeners to actually complete all of the tests. Some people simply forget to come back to complete their tests. Others forget which email address they registered. Others lose the evaluation URL. Finding ways to encourage completion from a large number of registered listeners is extremely important for upcoming years.

## 5. Acknowledgements

We would like to thank all of this year's participants and listeners. Thanks again to Richard Sproat for providing test sentences and to Brian Langner for maintenance of the evaluation database. Thanks also to Keiichi Tokuda, Simon King, Michael Picheny, and Toshio Hirai acting as the organizing committee. A specific thank you goes to Toshio Hirai and Satoshi Nakamura at ATR for arranging the release of their substantial speech database.

This work was funded in part by the US NSF grant (00205731) "ITR Prosody Generation for Child Oriented Speech Synthesis". Opinions expressed in this paper do not necessarily reflect those of NSF.

## 6. References

- [1] Black, A. and Tokuda, K., "The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common databases," in Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.
- [2] Kominek, J. and Black, A., "The CMU ARCTIC Speech Databases," SSW5, 2004, Pittsburgh, PA.
- [3] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S., "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Spain, May 2002, pp. 147-152.
- [4] Bennett, C. L., "Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005," in Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.
- [5] MacKenzie, I. S., "Within-subjects vs. Between-subjects Designs: Which to Use?" <http://www.yorku.ca/mack/RN-Counterbalancing.html>, 2002.
- [6] CCITT "Absolute category rating (ACR) method for subjective testing of digital processors," Red Book, 1984, Vol. V, (Annex A to Suppl. 14).
- [7] House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D., "Psychoacoustic speech tests: A Modified Rhyme Test," Tech. Doc. Rept. ESD-TDR-63-403, U.S. Air Force Systems Command, Hanscom Field, Electronics Systems Division, 1963.
- [8] Benoit, C. and Grice, M., "The SUS test: a method for the assessment of text-to-speech intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, Vol. 18, 1996, pp 381-392.
- [9] Levitt, S. D. and Dubner, S. J., *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*, William Morrow, New York, 2005.