Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005

Christina L. Bennett

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

Abstract

The Blizzard Challenge 2005 was a large scale international evaluation of various corpus-based speech synthesis systems using common datasets. Six sites from around the world, both academic and industrial, participated in this evaluation, the first ever to compare voices built by different systems using the same data. Here we describe results of the evaluation and many of the observations and lessons discovered in carrying it out.

1. Introduction

Evaluation is a necessary component of every research area. In the field of speech synthesis, several testing methods have been commonly used; however, when conducting evaluations, it is rare to compare voices developed using different systems. Most evaluation has been within a single research group for diagnostic testing or speaker selection. While a cross-system evaluation is obviously of value, results are not easily comparable if the data used to build the voices comes from different sources. The Blizzard Challenge was conceived by Alan Black and Keiichi Tokuda to eliminate this problem by specifying speech corpora on which all participating systems build voices, allowing for a meaningful comparison between sites and techniques [1]. The Blizzard Challenge was hosted by Carnegie Mellon University and conducted from February through April, 2005.

2. Blizzard evaluation methods

2.1. Listener groups

There were three categories of listeners in the evaluation: group S denotes speech *s*ynthesis experts from each of the participating sites; group V denotes *v*olunteers who heard about the evaluation from a mailing list or word of mouth; group U denotes *u*ndergraduate native US English speaking listeners, who were solicited and paid to participate. The primary reason for separating listeners into these groups was to allow us to compare results from the different populations. We hope to determine whether synthesis experts' opinions of synthetic speech differ from non-experts, whether a naive listener group of native speakers is required, or whether random volunteers from the web can be used just as reliably.

For group S, each of the participating sites agreed to provide 10 speech synthesis experts to perform the tests. Currently 50 (target 60) S listeners have completed them. For group V, we advertised the evaluation on various mailing lists and message boards, as well as spreading the word to colleagues around the world. While 97 people have registered, only 60 have completed every test. As for group U, we have found that a $5 payment is often not motivation enough for a sizeable number of participants of the type we defined. We offered this group an additional $10 incentive for doing again (in order to collect reliability data). At present, 63 have registered, 58 have completed all of the tests once, but only 10 have taken the tests a second time.

2.2. Test design

The evaluation was conducted entirely online using a simple web interface. While controlled laboratory listening experiments are ideal in many ways, allowing experimenters to maintain a consistent testing environment, there are other factors that favor a fully online configuration. With a web-based evaluation, listeners from diverse parts of the world can participate and complete tests at their own convenience, rather than committing to physically appear in a specific lab at a designated time. Access to a much larger population of listeners is the foremost reason for an online evaluation. Because of the international nature of this evaluation, we wanted to allow all parties who have submitted systems to be able to participate in the listening experiments.

The evaluation was composed of five separate tests, each of a different genre. Four voices from the CMU Arctic database [2] were used, *bdl*, *slt*, *rms*, and *clb*. For a complete description of the genres and voices, see [1]. Each test contained 20 samples, presented individually, preventing listeners from going back to alter previous responses once submitted. While each of the five tests could be done in any sequence, the 20 sentences in each test were presented in order. Listeners were given simple instructions for completing each of the tests. The main test page tracked the listener's progress, should they need to return at a later time to finish the tests. The five tests are estimated to have taken listeners roughly a total of 40 minutes to complete, with the most time spent on test 5, described below. Upon completing all five tests, listeners were asked to submit a feedback questionnaire containing demographic questions as well as voice preferences and open-ended comment sections. None of the feedback responses were strictly required.

2.3. Test ordering

Upon registration, each listener was assigned to one of ten groups, which determined system and voice ordering throughout the sentences and tests in the evaluation. Each listener heard all the voices but only one voice per test. Each test contained examples from all of the systems. For example, group A listeners would hear the *bdl* voice for tests 1 and 5, *slt* for test 2, *rms* for test 3, and *clb* for test 4, but group B listeners heard *slt* in tests 1 and 5, etc. Similarly, the ordering of systems was varied across tests and across groups.

Listeners were assigned to groups in batches of 10, in order to ensure enough listeners in each category so that consistency across listeners could be examined as well.

2.4. Test types

Tests 1 through 3 were Mean Opinion Score (MOS) tests [6], where the listener makes a judgment about the quality of a particular sample by assigning it a score of 1 to 5. The remaining two "type-in" tests required the listener to enter the words they heard into a textbox. The first of these tests was a Modified Rhyme Test (MRT) [4], where the same carrier phrase (i.e. "Now we will say __ again") is used for each example with a different word filling in the blank each time. Each MRT test word is confusable with at least five other words. The final test used Semantically Unpredictable Sentences (SUS) [5]. These sentences followed a simple grammatical structure (det-adj-noun-verb-det-adj-noun), but the words together were semantically nonsensical. The purpose of such a test is to test the intelligibility of speech in a controlled seven-word phrase for which listeners cannot use their higher-level knowledge to predict from semantic context the words spoken.

3. Results and discussion

To preserve participants' anonymity, the letters A through F are used to denote the systems, and X denotes a real speech reference condition of examples recorded by the voice talent. Reference condition X was tested alongside the other systems in the evaluation. Results across listener types will be compared to natural speech.

In Table 1, we list and rank the overall system scores on MOS and type-in tests, for each listener group. The average score for each system across the three MOS tests is listed first, followed by the overall word error rate (WER, the percentage of words that had errors) for the two intelligibility ("type-in") tests. No synthetic speech system approached the performance of natural speech, but relative performance of the TTS systems, in terms of rank on each test type across listener groups, was fairly consistent.

S Listeners		V Listeners		U Listeners	
MOS	*WER*	*MOS*	*WER*	*MOS*	*WER*
X - 4.76	X - 8.5	X - 4.41	X - 10.3	X - 4.58	X - 7.3
D - 3.19	D - 14.7	D - 3.02	D - 17.1	D - 3.06	D - 16.3
E - 3.11	B - 15.0	E - 2.83	A - 19.7	E - 2.83	A - 19.3
C - 2.91	A - 17.4	B - 2.66	B - 20.3	B - 2.67	B - 19.6
B - 2.88	E - 20.6	C - 2.48	E - 25.0	C - 2.42	E - 21.7
F - 2.15	C - 22.5	F - 2.07	C - 25.6	A - 2.00	C - 22.8
A - 2.07	F - 32.7	A - 1.98	F - 41.8	F - 1.98	F - 35.2

Table 1: System ranks by listener group and test type. MOS ratings are listed as well as WER for the type-in tests.

Another important observation is the similarity of listener groups V and U MOS results compared to group S. It is interesting to note that the MOS of V and U are extremely similar, whereas group S ratings tend to be slightly higher. WER is very similar across all three groups, with slightly more errors in the V group, which had a substantial number of non-native listeners. Recall that U listeners were restricted to native US English speakers; while group S included some non-natives, several opted not to perform the type-in tests due to this language barrier. Listener differences are discussed further in Section 4.1.

Listener group S				
Votes	*all - MOS*	*natural-MOS*	*all - WER*	*natural-WER*
slt - 20	rms - 3.233	bdl - 4.827	rms - 10.5	rms - 3.2
rms - 17	clb - 3.154	rms - 4.809	clb - 16.0	clb - 9.3
bdl - 8	slt - 2.994	slt - 4.738	slt - 20.8	bdl - 9.4
clb - 1	bdl - 2.941	clb - 4.690	bdl - 22.7	slt - 11.3
Listener group V				
Votes	*all - MOS*	*natural-MOS*	*all - WER*	*natural-WER*
slt - 23	clb - 2.946	rms - 4.568	rms - 14.0	rms - 3.8
rms - 13	rms - 2.894	clb - 4.404	clb - 17.1	bdl - 12.0
bdl - 9	slt - 2.884	bdl - 4.382	slt - 25.2	slt - 12.0
clb - 1	bdl - 2.635	slt - 4.296	bdl - 29.3	clb - 13.1
Listener group U				
Votes	*all - MOS*	*natural-MOS*	*all - WER*	*natural-WER*
slt - 26	clb - 2.987	slt - 4.611	clb - 11.9	slt - 5.9
rms - 19	slt - 2.930	clb - 4.587	slt - 17.5	clb - 5.9
bdl - 6	rms - 2.873	rms - 4.584	rms - 17.6	rms - 8.8
clb - 2	bdl - 2.678	bdl - 4.551	bdl - 28.7	bdl - 9.1

Table 2: Voice rankings based on exit poll votes (col. 1), overall results (cols. 2 and 4), and results on natural examples only (cols. 3 and 5), for each listener group.

Tables 2 and 3 compare the four voices used in the evaluation by listener group. In Table 2, voices are ranked by several measures: preferred voice chosen by listeners in an exit poll, MOS for all samples together, MOS on only the natural recordings, and likewise for WER on all samples and on natural speech. Exit poll results for all three listener groups are very consistent. The *slt* voice is preferred, with the *rms* voice a close second. However, this preference does not correspond to the MOS and WER results. In groups S and V, the *rms* voice outperforms the others, with six first place rankings, and two second; however the *slt* voice only manages to rank third (six times) and fourth (twice). On the other hand, the *clb* voice was not well liked by listeners, but its scores were reasonably good for these listener groups (1 first place, 5 second place). For listener group U, the *slt* preference seems well founded, ranking consistently in first or second place. Again the *clb* voice does well, sharing the top spots equally for U listeners, yet it only received 2 votes in the exit poll.

Table 3 further demonstrates this disparity but also highlights similarities and differences in the different systems' performances on each voice. The consistently better scores for *rms* in listener group S is elaborated here. Each system (except F which did not use *rms*) had its best WER score from this voice. Listener group U again in general shows better scores for the two female voices, *slt* and *clb*. In particular, *clb* was most often the best performing voice across systems for these listeners. Of the systems that included *clb*, only system B had better results on another voice for both MOS and WER. As for listener group V, systems' top ranking voices were spread among three (*slt*, *clb*, and to a lesser extent *rms*). The *slt* voice often gives a system its best MOS result, whereas a system's best WER often comes from *clb*. Only system B consistently had top scores from the same voices across listener groups. System F also had consistent results, but with only two voices, one of which was generally an underperformer for all systems.

S	***bdl***		***slt***		***rms***		***clb***	
	MOS	*WER*	*MOS*	*WER*	*MOS*	*WER*	*MOS*	*WER*
X	**4.827**	9.38	4.738	11.33	4.809	**3.19**	4.690	9.33
A	2.044	21.21	1.722	17.04	**2.333**	**11.75**	2.257	19.17
B	2.903	21.09	**3.046**	21.81	2.974	**8.84**	2.437	14.34
C	2.545	29.23	2.968	24.24	**3.196**	**17.40**	2.954	19.20
D	3.252	18.07	2.955	17.04	**3.324**	**8.27**	3.25	15.44
E	**3.325**	26.56	3.043	22.73	2.791	**15.49**	3.207	18.54
F	1.8	33.16	**2.492**	**32.33**	n/a	n/a	n/a	n/a
V	***bdl***		***slt***		***rms***		***clb***	
	MOS	*WER*	*MOS*	*WER*	*MOS*	*WER*	*MOS*	*WER*
X	4.382	12.02	4.296	12.16	**4.568**	**3.78**	4.404	13.10
A	1.868	24.38	1.702	22.96	2.022	15.31	**2.278**	**14.72**
B	2.688	29.61	**2.852**	26.27	2.771	**12.12**	2.314	22.57
C	1.980	38.43	**2.746**	25.48	2.638	21.67	2.575	**18.27**
D	2.974	24.68	3.041	20.21	**3.103**	**11.37**	2.988	13.47
E	2.812	31.19	**3.016**	30.31	2.448	21.33	2.963	**18.73**
F	1.727	43.93	**2.489**	**40.39**	n/a	n/a	n/a	n/a
U	***bdl***		***slt***		***rms***		***clb***	
	MOS	*WER*	*MOS*	*WER*	*MOS*	*WER*	*MOS*	*WER*
X	4.551	9.09	**4.611**	**5.92**	4.584	8.82	4.587	5.95
A	1.947	28.43	1.914	16.93	2.013	17.86	**2.116**	**11.39**
B	2.592	27.30	**2.922**	19.31	2.700	**15.20**	2.489	17.87
C	2.063	35.60	2.619	19.64	2.503	25.46	**2.549**	**11.25**
D	**3.232**	23.46	3.037	13.33	2.961	16.02	2.965	**12.01**
E	2.872	32.20	2.832	19.20	2.626	24.03	**2.946**	**11.07**
F	1.550	42.61	**2.528**	**27.38**	n/a	n/a	n/a	n/a

Table 3: Performance of each system on each voice; the best MOS and WER scores for each system are marked in bold.

WER performance was usually substantially better on MRT than on SUS, as expected. Statistical tests have not yet been fully analyzed; we plan to explore these areas further when the data are more complete.

4. Lessons learned

4.1. Lessons about listeners

As anyone who has ever done a study involving human participants knows, the most uncontrollable factors are the people themselves. We found there to be many issues stemming from these uncontrollable human factors, some of which were expected and many others of which were not.

A number of responses were excluded from the results presented in the previous section. Reasons for exclusion were: an incomplete test (complete tests from same listener were included, but partial tests excluded); failure to follow directions (e.g. wrote comments instead of the words spoken in type-in tests); inability to respond in type-in tests (because non-native); or "unusable" responses for any of the following reasons:

- lack of effort in type-in responses (e.g. "don't know")
- inappropriate responses (e.g. accidentally typed a previous sentence)
- scores extremely contrary to expectation (e.g. natural speech examples scored *very* low relative to others).

Eight people's results were found to have some "unusable" portion. While we made an effort to exclude the most obvious cases above, our exclusion of responses was conservative. With more examination of the data, likely more will be excluded from future calculations. One reason for being conservative in removing problematic users' responses entirely was that the effects of a "bad" (not serious) listener would be evenly distributed among systems because of the design of the tests; however, it should be noted that the effects may not be distributed evenly among the *voices*. If for example, there was an extreme lack of effort in the type-in tests, or simply an extremely high number of spelling errors, this would affect only the voice used in that particular test (especially likely for SUS).

There are a number of issues with the type-in responses since these were open, unrestricted inputs. While many of the problems can be resolved automatically (e.g. stripping punctuation, regularizing capitalization, etc.), several of the type-in responses require some level of manual correction. Unfortunately several homophones were present in the test sentences; we considered using a homophone list, but there were just as many (or more) obvious misspellings/typos as legitimately alternate words (homophones). In addition, since one of the participant sites is based in the UK, there would have to be separate homophone lists based on the dialect of English most often heard by a particular listener. For example, we noted after testing began, that the word 'bean' had been included in one of the MRT sentences. For a UK speaker, this could be either the word 'bean' (as we anticipated) or the word 'been'. In contrast, for some dialects of US English, 'been' has the same pronunciation as 'bin'; fortunately neither of those words was included in the task. The list of homophones clearly continues to grow with more responses, and thus is difficult to maintain automatically. At least one listener noticed that there were homophones (e.g. 'dug' vs. 'Doug') and listed both spellings separated by a slash (e.g. 'dug/Doug'); thus simple removal of punctuation introduces an error into what should have been a correct response. Other listeners actually included comments in their type-in responses, say within brackets after the true response. Additionally, in the MRT test where a carrier phrase is used, some listeners stopped typing the carrier phrase and simply entered the changing word.

Group V listeners were of course the most variable of the three populations. They were clearly the least motivated to complete the tests since the incentive for them was least, and many who registered never completed the tests. Responses from V listeners also seem to be much more varied and inconsistent, though a detailed analysis has not yet been undertaken. For this group it is important to closely analyze the responses received in order to determine whether or not they are serious. Many V listeners are non-native speakers, but since the question about native language is not given until after the tests have been completed and many in this category do not complete all the tests, it may be difficult to separately analyze non-natives' responses. Even when V listeners complete all of the tests, they are less likely to submit a full exit questionnaire.

Group U listeners are a more homogeneous population since they are specifically solicited based on demographic information (i.e. native US English speakers attending college). While these characteristics can be controlled, the seriousness of these listeners when taking the tests cannot. In this population, the listeners are often primarily motivated by the payment rather than by helping science and only sporadically answered the exit feedback questionnaire.

Group S, speech synthesis experts, were unsurprisingly the least problematic population. They have motivation to complete the task and to do so in a conscientious manner; however, there were still several who registered but never completed the tests. The MOS ratings of S listeners were on the whole higher than the other populations, and they were also most likely to give thorough feedback.

4.2. Lessons about test design

In general, listeners reported satisfaction with the design of the tests; nevertheless, some issues arose.

The most frequent dissatisfaction among listeners had to do with the scale used in the MOS tests. The majority of people who mentioned the scale said that they would have liked to have been given examples of some of the best synthesized samples and some of the worst, in order to calibrate for the 1 to 5 scale. Indeed, a brief familiarization phase is standard practice for MOS testing; however, we feel that our test ordering schema will counterbalance the effects of learning a user-defined scale over the first few samples.

Speech experts occasionally also commented on the MOS scale; however, this group was more likely to suggest having multiple scales for different dimensions such as naturalness and intelligibility. For this evaluation, we purposefully chose to use a single scale since lay-people would not understand the meaning of such dimensions and defining them sufficiently can be challenging (see [3]). We tested intelligibility directly through the "type-in" tests.

Other problems faced by listeners had to do with the type-in tests, particularly SUS. Some listeners expressed surprise at the nonsensical nature of these sentences, suggesting that they should have been forewarned. Another problem noted particularly by V and U listeners was that the sentences were too long to remember and the words too unusual, making a vocabulary/spelling test from what should be a fairly simple, stress-free exercise.

Certain users also faced an unforeseen problem with their own audio player / web browser setup. Though several different setups had been tested, we did not exhaustively test all combinations. For one embedded media player, listeners were taken to a new page, forcing them to hit the 'back' button on the browser in order to enter their responses. This introduced an unwanted memory component to the test.

In this evaluation, we opted to include natural speech examples recorded from the voice talent. We were able to do this since we collected the databases locally and had access to each of the original speakers. In future evaluations it may not be possible to elicit the set of test sentences directly from the original voice talent for whatever databases are employed. We found however that having this set was a valuable resource for determining listener seriousness. We were able to compare scores of different listeners at a glance by noting whether or not their scores were relatively high for the natural examples. In at least one case, this quick comparison allowed us to catch a listener who consistently scored the natural examples lowest, and upon closer inspection, exhibited a roughly opposite distribution of scores to the other listeners. It is unclear however whether this was an accidental inversion of the scale or a 'malicious' listener.

The virtues of natural speech references aside, we note that evaluating them can be difficult, particularly because of differences in delivery style. Some of the voice talent used a very natural prosody, whereas others gave a more flat delivery, in order to be consistent with that provided previously for the purpose of building synthetic speech. Listeners often seemed to be influenced by these prosodic factors, which is of course reasonable, but may have lowered the scores of natural examples from certain speakers.

After completing the tests, listeners were asked to select which of the four voices they most preferred. In order to do so, they were presented with a natural sample from each of the speakers. Comments from listeners suggested that the order of presentation of these natural samples influenced their choices. Listeners also suggested that we include a question about which voice was *disliked* most. These comments seem to imply that a ranking or scoring system would be better than a simple preference question. Another option would be to ask more detailed questions about qualities they liked or disliked in each voice.

5. Acknowledgements

Thanks to all the participants and listeners and especially to Alan Black and Keiichi Tokuda for initiating the challenge. Thanks to Ann Syrdal of AT&T Labs Research for her input in preparing this report and also to Dan Jurafsky of Stanford University and Chris Brew of Ohio State University for helping us find willing undergraduate listeners.

This work was funded in part by NSF grant (0219687) "ITR/CIS Evaluation and Personalization of Synthetic Voices". 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." http://www.festvox.org/blizzard

[2] Kominek, J. and Black, A., "The CMU ARCTIC Speech Databases," SSW5, 2004, Pittsburgh, PA.

[3] Vazquez-Alvarez, V. and Huckvale, M. "The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems," ICSLP, 2002, Denver, CO, 329-332.

[4] 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.

[5] 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.

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