The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets

Alan W Black[1] *and Keiichi Tokuda*[2]

[1]Language Technologies Institute, Carnegie Mellon Unversity, Pittsburgh, PA, USA
[2]Dept. of Computer Science & Engineering, Nagoya Institute of Technology, Nagoya, JAPAN
awb@cs.cmu.edu, tokuda@ics.nitech.ac.jp

Abstract

In order to better understand different speech synthesis techniques on a common dataset, we devised a challenge that will help us better compare research techniques in building corpus-based speech synthesizers. In 2004, we released the first two 1200-utterance single-speaker databases from the CMU ARCTIC speech databases, and challenged current groups working in speech synthesis around the world to build their best voices from these databases. In January of 2005, we released two further databases and a set of 50 utterance texts from each of five genres and asked the participants to synthesize these utterances. Their resulting synthesized utterances were then presented to three groups of listeners: speech experts, volunteers, and US English-speaking undergraduates. This paper summarizes the purpose, design, and whole process of the challenge.

1. Background

With a view to allowing closer comparison of corpus-based techniques, from labeling, pruning, join costs, signal processing techniques, and others, we devised a challenge for participants to use the same databases to synthesize utterances from a small number of genres. An organized evaluation, based on listening tests, was then carried out to try to rank the systems and help identify the effectiveness of the techniques.

The sister field of speech recognition has clearly benefited from the availability of common datasets in order to provide valid comparisons between systems [1]. These evaluations concentrate efforts in the speech recognition fields, particularly through the 1990s with DARPA workshops where NIST (and others) devised standardized tests for speech recognition. It is clear that these standardized tests and widely available datasets allowed speech recognition results to be more easily compared and more importantly cause the core technology to improve. Although today many may criticize a naive word error metric as a sole accuracy measure for speech recognition systems, few would complain that it has not contributed to drastic improvement in the utility of speech recognition as a viable technology.

Speech synthesis has not been as lucky in having a well-defined evaluation metric, nor has it had a well-funded centralized community that could be targeted to the same task. With the rise of general corpus-based speech synthesis over the last ten years, we have moved from a domain where new synthetic voices could only be built with many man-years of effort from highly skilled researchers. Such systems were tuned to the particular data sets being used, thus comparisons of techniques such as labeling and signal processing could only be done within the research group that originally developed the dataset. Such tying of databases to particular systems made it hard to genuinely compare techniques since the quality of the original recorded voice itself contributed greatly to the resulting synthetic voice quality.

2. A common dataset

The key aspect of the Blizzard Challenge [2] is a common dataset shared between participants. Removing the variability of the data itself allows for a much closer comparison of the voices generated from the data. There are number of issues which must be answered before we can provide such a set. The issues include: what size and shape should the data be; who will collect and release it; what additional data may also be used.

As part of Carnegie Mellon's effort to make speech synthesis research more accessible, a set of single-speaker phonetically-balanced databases have already been constructed. The CMU ARCTIC databases [3] were designed with speech synthesis in mind. They consist of around 1200 phonetically-balanced sentences selected from out-of-copyright texts. The initial release in 2003 consisted of four speakers: one male US English (**bdl**), one female US English (**slt**), one male Scottish English (**awb**) and one male Canadian English (**jmk**). Because the last two speakers have non-standard dialects, it was decided not to include them in the Challenge. In addition, in order to properly test the participants' capabilities with voice building, we decided to release two further databases (male (**rms**) and female (**clb**)), at the time of the challenge to test not just how people built voices, but how quickly they could do so. Thus **bdl** and **slt** were available from the original call for Blizzard participants (June 2004), whereas **rms** and **clb** were released with the set of test utterance texts (Jan 15th 2005). **rms** had in fact been recorded in the summer of 2004, while the **clb** database was recorded the week immediately prior to the final release.

The idea of a common dataset has been discussed in the synthesis community, but no acceptable set has appeared. The TIMIT (452) set of sentences is a potential, but many have failed to successfully use such a set: although University of Edinburgh's CSTR has released a single-speaker version of TIMIT [4], we know of no one who has successfully used it as it is small by common standards. The Boston University FM Radio corpus [5] is another possibility and has been used for a number of prosody experiments, but it is not phonetically well balanced. In Japan, the ATR 503 sentence sets [6] have long been used as a common dataset among Japanese researchers, but there is no easy way for people outside Japan to gain access to these even if we decided to test our techniques in Japanese. Most commercial synthesis companies have a number of large single-speaker high-quality databases. However, even though older versions often become unused, it is hard for a company to consider releasing these (potentially to commercial competitors), even if only for research purposes. It is not just the data that is valuable, but it is the content and design of the database

that many companies feel proprietary. Furthermore, data costs money and few wish to let others look at it. Thus, at least for the initial stages of the challenge, we used the freely available single-speaker CMU ARCTIC databases.

The CMU ARCTIC voices consist of around 1200 phonetically-balanced utterances. The data was selected from a number of novels from Project Gutenberg [7]. These sentences were chosen to be easy to read, restricting their length from 5 to 15 words, and that all the words were already in CMUDICT. The sentences that fit this "niceness" criteria were then synthesized as strings of phonemes and we greedily selected sentences with the maximum diphone coverage. This selection process was done twice, thus the 1200 utterance set consists of an A and B set (each around 600 utterances) each of which is itself phonetically balanced. Sentences are typically easy to read, and in prose style, though they sometimes appear archaic as they are all extracted from out-of-copyright books. Typical sentences are

- *Author of the danger trail, Philip Steels, etc.*
- *I was the only one who remained sitting.*
- *They were three hundred yards apart.*

These were recorded at 16KHz, 16bit, in studio quality conditions. An electroglottograph (EGG) track is also included. In addition to the raw data, a basic build using the publicly available FestVox tools [8] was included. This includes phonetic labels generated using the CMUDICT [9] and forced alignment using speaker specific HMM acoustic models trained using SphinxTrain [10]. The resulting labels were provided in a Festival Utterance structure [11], a structure that is currently used by many in the field. Although participants need not use this extra information, we made it available.

Although we feel the CMU ARCTIC data is a reasonable starting point, some participants' systems are designed for dealing with much larger datasets (and one that uses much less). There have been a number of successful concatenative synthesis systems that use much much larger databases, and take advantage of that variability. By specifying a database smaller than they are used to, it may unfairly degrade their system in such a test. Also commercial grade corpus-based synthesizers typically include prompts from various domains. On the other hand, the CMU ARCTIC's data is predominantly from novels making it less than ideal, though in our testing we do test "in-domain" sentences (from novels) as well as other genres.

3. An evaluation methodology

Although the over all goals of the challenge are clear: finding better synthesis techniques by comparing systems on the same data, those working in speech synthesis know "better" is not "better" for everyone all of the time. As it is not clear what the best evaluation technique is, we decided to pick several simple ones, and look at the results to determine which tests were sufficient, and/or more reliable.

First we decided to choose 5 different genres of text to be synthesized. We did not want this to be a text analysis exercise, so the text itself would be relatively simple (without many numbers, symbols etc). We chose to have two basic types of tests. The first three genres were for simple mean opinion score (MOS) tests where the listeners are asked to rate their views of the synthesized texts on a score of 1 to 5. The final two genres involved the listeners typing in what they heard, in order to specifically address intelligibility in the task and not just personal preference.

The five genres we chose were:

novels taking text from stories, specifically the same stories (but different text) from which the CMU ARCTIC databases were built, thus this could be considered an "in-domain" task.

- *Joe Garland lives like a good fellow.*
- *But we made no collections of eggs.*

news taking text from standard press-wire news stories.

- *The two countries agreed to resolve any conflict through diplomacy and avoid the use of force, the agency Interfax said.*

conversation in an attempt to deal more with the sort of speech one would expect in a spoken dialog system, we took examples from the *human* side of a spoken dialog system. Although this sometimes produced some unusual utterances, it did provide utterances that were quite different from standard TTS utterances.

- *Okay I would like to go to Miami, Florida.*
- *Yeah I guess it will and something downtown please.*

phonetically confusable sentences following the DRT (and MRT) [12], we constructed carrier phrases that contained phonetically confusable words. The isolated word tests of the original DRT/MRT word lists are not appropriate for unit selection based synthesis where different units will likely be selected in the isolated word case than in the intra-sentential case. Putting the confusable words in carrier phrases seems a more reasonable test.

- *Now we will say cold again.*
- *Now we will say pace again.*

semantically unpredictable sentences (SUS), following [13] we constructed a simple grammatical template, *det adj noun verb det adj noun*, and randomly generated words of medium frequency. The resulting sentences are hard to understand and remember even when spoken by human speakers.

- *The unsure steaks overcame the zippy rudder.*
- *The dank geniuses woke the humane emptiness.*

Prof. Richard Sproat of UIUC generated the actual example sentences, as we wished some level of independence from the organizers who's teams were also participating in the challenge.

50 sentences for each genre were distributed to the participants on 15th January 2005. In addition to the sentences themselves, we also released Festival-style utterance structures for each sentence.

4. The participants

Our initial discussions found around 8 potential participants. However, when the time came for the tests, we found that some could no longer participate. One, a commercial company, had been bought and no longer wished to take part, and another commercial company was worried about the publication of results. It should be remembered that for commercial companies to take part in an exercise like this does involve a substantial risk. In the case where their system is particularly bad, even if they can "blame" the inadequate data, it can be detrimental to their bottom line.

Thus we expected participants to be mainly research groups, which was true, but they include one large industry research group. It was, somewhat tongue in cheek, pointed out to

us that commercial systems do not need to be evaluated as each company's system is **always** the best system.

In all we have 6 participants, spanning 3 continents, including both universities and one commercial company. Some of these groups already had a close working relationship (CMU, Edinburgh and NITECH) who had already shared engines, data and techniques, but the participants also included other groups who had not had a history of working closely together.

5. The Challenge

On January 15th 2005 we released the test sentences to the already registered participants. As described in 2, two of the voice corpora were released well before the Challenge (**slt** and **bdl**), and a further two new voices (**rms** and **clb**) were released at the same time as the test sentences.

Although at first we were concerned that participants may tune their systems to the particular sentences, we decided not to institute any explicit safeguard against this since it would be too restrictive and we trusted the participants not to do this.

We allowed the groups one week to build their new voices and synthesize the 250 test sentences. However, we allowed the return deadline to slip for a second week. There was at least one additional group who did significant work during that time but did not have their algorithms mature enough to complete the task.

Once the Challenge was opened, a number of other sites indicated their interest in taking part but had missed the original call for participation. In the future, we hope to overcome these missed opportunities, by better advertising and also by being already better known.

An additional "participant" was added to the mix. Because the original speakers of the four databases were available at Carnegie Mellon University, we recorded the 250 test sentences from the human speakers. Of course we expected such a "team" to do well in such a competition, but their participation was intended to allow us to find out how far our synthesizers are from natural speech.

6. The tests

Due to the number of expected participants (and the number of different files that would need to be listened to), from the start we decided to carry out all listening tests across the web.

There are distinct disadvantages to using the web for such tests. The environments where the listeners are in may be very different. Although we asked the listeners to listen in a quiet room, there is no guarantee that they would. Network congestion can make downloading of waveform files slow and even intermittent. Such waveforms may appear to have join problems when the network congestion is at fault (there is evidence from user feedback that this did actually happen). Also we would have no control over the listeners so they may become disinterested and not finish tests or worse just fill in random values.

Nevertheless, there are also advantages to web delivery. We had participants from many continents and listeners from even more. Finding our target 200-300 listeners is hard and we wanted to easily get as many as possible. There was also the question of where web-based listening tests are feasible for large scale synthesis evaluation and this would give us an opportunity to find out. Therefore in spite of the expected noise that would be introduced from web tests, we felt that it was worth the advantages of accessing more people.

The tests themselves (which can be carried out by visiting the Blizzard homepage [2]) consist of five tests, one for each genre. Each test consists of listening to a total of 20 different utterances synthesized by one voice but by randomly ordered different synthesizers. Only 20 sentences randomly chosen for each listener from the 50 submitted were used in the listening tests, as listening to all 50 would require too many resources. Our target was to have 10 listeners for each sample. Each sample was presented singly and had to be rated before proceeding to the next sample. This was done to stop people changing their minds, though at the cost that people's scaling was settling during the first few examples.

The tests could be suspended and continued at a later date, though we encouraged people to do the tests in one sitting. To complete all five tests took around 30-45 minutes.

7. The listeners

The issues of who should evaluate speech synthesis systems is an open question. Experts in the field have detailed knowledge and will listen for intonation, join and phonetic errors, or artifacts introduced by signal processing. The real user (whoever they may be) may care about the accent or whether it sounds like someone they know (or not). Real users are unlikely to be able to identify why something sounds strange. Therefore lots of different listeners are required to be able to achieve stable results.

We decided to address three different listening groups, in part to be able to compare these groups' views with each other, in addition to evaluating the synthesis performance itself. The three groups were:

speech experts we made it a requirement that all participants provide 10 of their local speech experts to do listening tests. We see this group as the most knowledgeable about speech synthesis, and also will represent the group that synthesis developers will most likely turn to when evaluating their systems. We did allow groups to evaluate their own systems too.

volunteers these were in early drafts of the Challenge sometimes referred to as "random" users. This group was those available on the web who were willing to take part. This did include some speech experts and even some speech synthesis experts. We advertised for this group though mailing lists and web pages.

US undergraduates the third group consisted of US English speaking undergraduates. Unlike the other two groups we paid these listeners. Our goals were to find a consistent set of listeners who were paid to do what is not a very interesting task.

These three groups would have very different goals in carrying out the tests, and we want to find out if each group was internally consistent as well.

As it turned out the first group were the most eager to carry out the tests and the most conscientious in completing them. The second group were more "random." Some did well, and some never completed. The third group proved to be surprisingly hard to convince to do this. We offered a $5 Amazon.com gift token for completion of all five tests, but a further $10 gift token if they did a second 5 test session. Although these fees are consistent with what we pay undergrads to take part in other experiments, it still proved hard to attain our goal of 100 undergrads.

8. The unknown issues

Organizing such an effort will always have its unknowns. One specific issue that came up is how much extra data should be

allowed in building a voice. When the challenge was first devised, we expected both concatenative synthesis techniques and HMM-based techniques, and perhaps some interesting variations, e.g., diphones selected from the data.

One direction we did not initially consider is voice conversion techniques. One potential participant did construct voice conversion entries. This was not something we had considered and had already designed the listening tests to be MOS (1–5) tests rather than AB, ABX, and/or DMOS tests because that would have presented us with too many samples to find listeners for.

Although no one did this, in an extreme case, we could imagine a entry that uses nothing of the common dataset. Given the current evaluation paradigm, such an entry, if say one of the flagship voices from a large voice foundry, would probably do very well. Deciding on the line when entries are not using the data enough is going to be hard to define.

To validly include voice conversion entries, we would need to also include some test of similarity with the original speakers. In this case, we decided not to include this potential entry in our tests, but would like to consider such entries in the future since voice conversion is a perfectly valid technique for constructing synthetic voices.

9. The results

The results of the evaluation are discussed in [14]. In order to allow for teams who do not wish their identity to be revealed, we have decided not to explicitly reveal the participants names nor their ranking in the results. The final results are published with team letter names, while the teams know their own corresponding letter. Although they do not know the others, we suspect that the teams themselves will work out which is which.

10. Releasing the data

We now have same 250 utterances per speaker per system, plus the Team Studio natural recordings. 100 utterances of each entry have been listened to by some 10 or more listeners. This data is valuable to others in research for automatic measures for synthesis evaluation. Thus it is our intention, with the participants permission, to release this data to allow further study.

11. Suggestions for the future challenge

We hope to make this an annual event. Although there are many ways to extend the evaluation methodology, we think that drastic changes should not be made for the next challenge since it may confuse participants and makes it difficult to use the feedback obtained from the challenge in 2005. We may repeat the similar process at least three times.

Changes we should consider for the next challenge could be

larger database the database size could be larger since participants whose system designed for larger databases prefer to use larger ones. The problem is that providing freely available large databases is difficult as described in 2.

similarity measure the final target of corpus-based speech synthesis is to synthesize utterances identical with those uttered by the original speaker for any sentences. To evaluate similarity with the original speaker, we may conduct additional DMOS (degradation mean opinion score) tests, though it increases the cost of the evaluation.

12. Conclusions

The Blizzard Challenge has been a valuable exercise in building voices from a common dataset and has brought together different teams looking at a common goal. We see this Challenge as the start of a series of synthesis Challenges, and organization has already started for a Blizzard Challenge 2006.

13. Acknowledgments

We would like to thank, Stefanie Tomko, Brian Langner, Christina Bennett and Rich Stern for their substantial help in speech database collection.

Thanks go to Dan Jurafsky of Stanford University and Chris Brew of Ohio State University for helping us find willing undergraduate listeners. Also thanks go to Christina Bennett and Brian Langner who created and ran the evaluation experiments.

We would also like to thank Richard Sproat of UIUC for generating the test sentences, and the participants of the challenge itself for taking part. This work was partially supported by the US National Science Foundation under grant number 0219687 "Evaluation and Personalization of Synthetic Voices". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

14. References

[1] DARPA, "The DARPA speech recognition evaluation workshops," http://www.nist.gov/speech/publications/index.htm.

[2] A. Black and K. Tokuda, "The Blizzard Challenge 2005," 2005, http://festvox.org/blizzard/.

[3] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu_arctic/, 2003.

[4] CSTR, "CSTR US KED TIMIT," 2001, University of Edinburgh http://www.festvox.org/dbs/dbs_kdt.html.

[5] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA, Tech. Rep. ECS-95-001, 1995.

[6] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuwabara, "A large-scale japanese speech database," in *Proceedings of ICSLP 90*, 1990, pp. 1089–1092.

[7] M. Hart, "Project Gutenberg," 2000, http://promo.net/pg/.

[8] A. Black and K. Lenzo, "Building voices in the Festival Speech Synthesis System," 2000, http://festvox.org/bsv/.

[9] CMU, "Carnegie Mellon Pronuncing Dictionary," 1998, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[10] Carnegie Mellon University, "SphinxTrain: building acoustic models for CMU Sphinx," 2001, http://www.speech.cs.cmu.edu/SphinxTrain/.

[11] P. Taylor, A. Black, and R. Caley, "Hetrogeneous relation graphs as a mechanism for representing linguistic information," *Speech Communications*, vol. 33, pp. 153–174, 2001.

[12] J. Logan, B. Greene, and D. Pisoni, "Segmental intelligibility of synthetic speech produced by rule," *Journal of the Acoustical Society of America*, vol. 86(2), pp. 566–581, 1989.

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

[14] C. Bennett, "Large scale evaluation of corpus-based synthesizers: Results and lessons from the blizzard challenge 2005," in *submitted to Interspeech 2005*, Lisbon, Portugal, 2005.