The Blizzard Challenge 2007

Mark Fraser and Simon King

Centre for Speech Technology Research, University of Edinburgh
m.e.fraser@ed.ac.uk, Simon.King@ed.ac.uk

Abstract

In Blizzard 2007, the third Blizzard Challenge, participants were asked to build voices from a dataset, a defined subset and, following certain constraints, a subset of their choice. A set of test sentences was then released to be synthesised. An online evaluation of the submitted synthesised sentences focused on naturalness and intelligibility, and added new sections for degree of similarity to the original speaker, and similarity in terms of naturalness of pairs of sentences from different systems. We summarise this year's Blizzard Challenge and look ahead to possible designs for Blizzard 2008 in the light of participant and listener feedback.
Index Terms: Blizzard Challenge, speech synthesis, evaluation, listening test

1. Introduction

The Blizzard Challenge was conceived by Alan Black and Keiichi Tokuda with the aim of comparing research techniques in building corpus-based speech synthesisers [1]. In each annual Challenge, a speech database is released to registered participants to build synthetic voices. A set of test sentences is then released for participants to synthesise. A subset of the synthesised sentences are evaluated in listening tests. Blizzard 2005 and 2006 were organised and run by CMU. Blizzard 2007, run by the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, preserved the main features of the previous challenges and introduced new ones. For general details of Blizzard 2007, the rules of participation, a timeline, and (in due course) information on forthcoming Blizzard Challenges, see [2]. In this paper we summarise Blizzard 2007 – participants, voices to be built, evaluation design, results, and listener feedback - and consider possible designs for the next Blizzard Challenge.

2. Participants

The Blizzard Challenge 2005 [1, 3] had 6 participants and Blizzard 2006 had 14 [4]. In 2007, the number of entries increased again: 19 sites registered, 18 returned signed licences for the data and 16 submitted entries:

- CereProc Ltd, UK
- Carnegie Mellon University, USA
- CSTR, University of Edinburgh, UK
- DFKI GmbH, Germany
- HTS working group (Nagoya Institute of Technology, Nara Institute of Science and Technology, University of Edinburgh), Japan and UK
- iFlytek Research, P.R. China
- INESC-ID, Portugal
- IVO R&D, Poland
- mXac, Australia
- Nokia Research Center Beijing, P.R. China
- SVOX AG, Switzerland
- Toshiba, UK and Japan
- University College Dublin, Ireland
- Universitat Politecnica de Catalunya, Spain
- University of Science and Technology of China
- Voiceware Co. Ltd, Korea

See Tables 6 to 8 at the end of this paper for an overview of the systems' characteristics based on participants' responses to a questionnaire.

3. Voices to be built

The data for voice building was provided by ATR. Participants who had signed a user agreement were able to download about 8 hours of data which had been made available from the 16 hour ATR English Speech Corpus of an American English speaker. For further details on this corpus, including the number of sentences and phoneme coverage of the Blizzard 2007 subset, see [5].

Participants were asked to build three synthetic voices from the database, using the same method, software, external data, etc.

- Voice A: from the full dataset (about 8 hours)
- Voice B: from the ARCTIC subset [6] (about 1 hour)
- Voice C: from a subset of the data chosen by each participant, under the following conditions:
  - selection could only be based on the text (and not the speech, or any information such as labelling which had been derived with reference to the speech signal)
  - if the selection method required phonetic, prosodic, or any other type of labelling, this had to be derived from the text only
  - entire utterances had to be selected
  - the total duration of the utterances selected had to be no more than 2914 seconds (the duration of the ARCTIC subset) - participants were provided with a durations file to make this calculation
- if the provided database was used to train any parts of the system (e.g., a prosodic model or HMM parameters), then for voices B and C, the whole database could not be used to train those parts, but only the appropriate subset.

For full details of the Blizzard 2007 rules, e.g. use of external data, see [7].

4. Test Sentences

The 400 test sentences were from two sources:

- 300 held-out sentences from the ATR corpus [5] in the following genres:
  - Conversational (100)
  - News (100)
  - ARCTIC (100)

  For each genre the total number of sentences was double that distributed in 2006 in order to discourage manual intervention during synthesis.

- Sentences designed for intelligibility tests, newly generated by Richard Sproat:
  - Modified Rhyme Test (MRT) [8] - 50
  - Semantically Unpredictable Sentences (SUS) [9] - 50

All 16 participants submitted samples of voices A and B for evaluation; 11 submitted samples of voice C.

5. Evaluation

The evaluation was conducted online. We were very fortunate to receive materials from the previous organisers at CMU for the web interface and database. This enabled us to build on the design developed over the last two Blizzard challenges [1, 3, 4], and adapt certain features in response to previous listener and participant feedback. We also wanted to incorporate current research at CSTR on speech synthesis evaluation using Multi-dimensional Scaling (MDS) techniques [10]: section 2 of the evaluation was designed to tell us more about the evaluation itself and to try to identify which features the listeners were focusing on when making their responses with respect to the 'naturalness' of samples.

5.1. Listener types

Four listener types were used (letters in parenthesis show the identifiers used for each type in the distributed results):

- Speech experts (S). Participants were asked to recruit at least 10 speech expert listeners each, preferably native speakers.
- Paid US undergraduates, native speakers of US English, assumed to be aged about 18-25 (U). These were recruited by contacts in USA, and by advertising for US students studying in Edinburgh to do the evaluation in a supervised lab.
- Paid UK undergraduates, native speakers of British English, aged 18-25 (K). In order to to boost the number of completed evaluations and as a means of ensuring that paid subjects matched the constraints required for the paid listener groups, we introduced this new listener group. Evaluations with this listener group were conducted in supervised labs.
- Volunteers (R) - 'real people'.

5.2. Interface

The registration page for each listener type presented an overview of the purposes of the challenge and tasks. Since some listeners in previous evaluations had felt that *they* were being tested [4], it was made clear in the instructions that it was the listeners who were doing the testing, and the word 'test' was avoided elsewhere. On registration, in order to reduce the load of the task, as in Blizzard 2006 [4] listeners were assigned to hear only voices built with one of the datasets - A (full dataset), B (ARCTIC), or C (subset of the data chosen by each participant).

There were 5 sections in the evaluation. They could be done in any order, though the order presented was designed to take listeners from lighter tasks to more difficult ones and was intended to improve completion rates. Listeners were encouraged to do the evaluation in a single session, estimated at 45 minutes (possibly longer for non-natives), but the evaluation could be done in multiple sessions if desired. On completion of any section, or after logging in again, a progress page showed listeners how much they had completed. Detailed instructions for each section were only shown on the page with the first part of each section; subsequent parts had briefer instructions in order to achieve a simple layout and a focussed presentation of the task. Sections 1 and 2 were new for Blizzard 2007; the tasks in sections 3, 4 and 5 were very similar in design to those in some sections of Blizzard 2005 and 2006 but instructions were rewritten and the interface changed. Since media player issues (such as pop-up windows, or web browsers navigating to a new page when sound files were played) were reported in [4] as a primary cause of complaints by listeners, we used an embedded media player design.

5.3. Listener tasks

We will now look at the tasks in each section and how they were presented to the listener.

- Section 1: In each part, listeners could play a fixed set of 4 reference utterances from the original speaker (2 taken from ATR conversational data and 2 from ATR news data) and one synthetic sample. They were instructed to choose a response that represented how similar the synthetic voice sounded to the voice in the reference samples on a scale from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person]. This section was introduced primarily because statistical parametric synthesisers have the potential to sound like another speaker (e.g., if the models have been trained on speech data from other speakers and then adapted to the target speaker).
- Section 2: In each part, listeners heard pairs of different sentences - one sample from each of two of the participating systems, or, in the case of one system ordering for each dataset (see Section 5.4), two samples from the same system. Listeners were to ignore the meanings of the sentences and instead concentrate on how natural or unnatural each one sounded. They then chose whether, in their opinion, the two sentences were similar or different in terms of their overall naturalness. This section was designed to be analysed using Multidimensional Scaling (MDS) techniques [10]. Since this analysis is more complex, it was not available to participants until the time of the Blizzard 2007 workshop; see [11].
- Section 3: Mean Opinion Score (MOS), conversational domain. In each part of sections 3 and 4, listeners were presented with one utterance and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 [Completely Unnatural] to 5 [Completely Natural].
- Section 4: MOS, news domain.
- Section 5: Semantically Unpredictable Sentences (SUS) designed to test the intelligibility of the synthetic speech. As in previous years the structure of

the sentences was det-adj-noun-verb-det-adj-noun, although this was not explicitly stated to the listeners. Listeners heard one utterance in each part and typed in what they heard.

Some listeners in previous evaluations reported that they had trouble calibrating MOS scales until they had already listened to some examples and submitted scores, which they could not then go back and change [3]. The ordering of the different sections this year was intended to make listeners familiar with the range of synthesised examples, including comparison with samples of the original speaker and with samples from all systems in terms of their naturalness, before reaching sections where they had to give MOS scores for single samples based on how 'natural' they sounded. An alternative would have been to present the listeners with some calibration samples (chosen by the experiment designers) that represented the extremes of the scale. This was not done since any such calibration examples would themselves be chosen subjectively. The variety of system orderings used in the evaluation also helps diminish the statistical impact of any calibration effects in the early parts of the MOS tests.

In the SUS test, because of issues reported in previous challenges [3, 4], where listeners had complained that they should have been forewarned about the nonsensical nature of sentences, and that the difficulty of the sentences contributed to some listeners feeling their intelligence was being tested, the instructions explained that sentences were not intended to make sense, that some might be unintelligible, that they might include unusual words, but that listeners should enter all the words they heard. Listeners were asked to limit the number of times they played each sentence to the fewest possible. Since it was anticipated that some listeners might give up during this task, they were encouraged to complete all parts even if they found it difficult.

Several features of the evaluation design were intended to maximise completion rates. It was not too long, it could be done in stages, and the tasks were presented in order of perceived difficulty, though listeners could choose to do them in any order if they preferred. They were told how many sections to expect and could see their overall progress after completing a section or logging in again.

5.4. Listener groups and system orderings

We look now at the underlying experimental design. Following the Blizzard 2006 design, the number of listener groups for each voice was determined by the total number of systems which had submitted samples for that voice plus the original speaker, i.e. 17 for each of voice A and B, and 12 for voice C. System orderings were systematically varied by using a Latin Square design. For sections 1, 3 and 4, Voices A and B required order 17 Latin Squares and Voice C order 12. Section 5 required a variation: there were no recordings of the original speaker reading the SUS sentences, so there were 16 systems for voices A and B and 11 for voice C. Since we already had 17 listener groups for each of voice A and voice B, as required by sections 1-4, an extra row was added to an order 16 Latin square (for voices A and B) and to an order 11 Latin square for voice C, in order to provide enough system orderings for the 17 or 12 listener groups respectively. This additional row was taken from another Latin Square of the same order, i.e. no complete system ordering (a row of the square) was repeated. A consequence of this is that, in Section 5 only, each system appears in the same position (column of the square) twice, although the surrounding context differs.

Distinct Latin Squares were constructed for all voices and sections. The rows of the squares corresponded to the listener groups, the columns corresponded to the sentences. The symbol (a letter from A to Q) in each cell (i,j) represented the system that listener group i heard reading sentence j. For each listener group, each test had a different ordering. No system was in the same slot across the two MOS tests for any single listener group. The evaluation was designed to minimise possible ordering effects. The Latin Squares were as balanced as possible, but Latin Squares of odd order - required for some voices and sections due to the number of systems - cannot be perfectly balanced.

In section 2 each listener group would hear 17 (voices A and B) or 12 (voice C) of the total possible pairings of systems (including the original speaker), in both orderings. The pairs were of differing sentences. A Graeco-Latin square design was used in order to distribute the pairs across the listener groups so that each pair was unique and each system appeared once as the first and once as the last of a distinct pair in each row of the square (listener group); systems appear in first or last position for any slot once only. The same-system pairs are all in one row (listener group) because otherwise the other constraints cannot be satisfied. This was necessary, although it was admittedly confusing to some of the listeners in that listener group, who detected that all of the pairs sounded the same in terms of naturalness. This was not always the case though - some listeners responded that same-system pairs were different in terms of their overall naturalness.

5.5. Listening test sentence selection

The sentences for participants to synthesise were randomly selected from the held out ATR data. In order to select the sentences for use in the online evaluation, the conversational and news sets were re-shuffled and where there was no valid reason for exclusion the required number of sentences were simply taken in order from these shuffled sets. Criteria for exclusion included

- sentences with features that would be a test of text normalisation
- sentences containing foreign words
- sentences containing more than one sentence (e.g., question and statement)
- sentences that were clearly ambiguous in how they should be read

We also tried to select sentences of a similar length because some sections of the listening test involve pairs of sentences. A final check was made that, where applicable, the original speaker read each sentence well, without disfluencies or mispronunciations.

The MRT section of the listening test was dropped this year, partly because it is tedious for listeners, but mainly because space was needed for the MDS section: we desired the listening test to have a maximum duration of around 45 minutes for most listeners. We also dropped the MOS section based on ARCTIC sentences for this reason.

The sentences in section 2 (MDS) were all used in the MOS tests as well. In all but one case, the two sentences within a pair being compared were from the same genre - conversational or news. Sentences of similar length were used in all pairs. The reason for repeating the MDS sentences in the MOS tests was that it would allow us to compare MOS scores and position in MDS space.

5.6. Listener numbers

The listener responses used for the distributed results were extracted from the database on 8th June 2007 at 04:43 BST. The online evaluation had been running for just over six weeks. 489 listeners had registered, of whom:

- 306 completed all sections of the evaluations
- 97 completed some of the evaluation
- 86 entered no response at all.

See Table 9 for a detailed breakdown of evaluation completion rates for each listener type.

6. Results

A full description of the statistical analysis of the listening test results that were performed by the organisers and distributed to participants, is provided in [11].

In Blizzard 2006, statistics were presented for two conditions: "strict" (using responses only from listeners who completed the whole listening test) and "lax" (using responses from all listeners, but discarding partially-completed sections). The two sets of statistics generally agree. It appears that listeners who do not complete the entire test still provide consistent responses. In other words, we do not believe they give up because they are having difficulty with the task. Therefore, in 2007, we used all listener responses to compute the summary statistics. Since the listening test design is based on Latin Squares, and therefore each individual subject only hears a small part of the whole dataset, this does not unbalance the design. Note that, in order to keep the design as balanced as possible overall, we attempted to get equal numbers of listeners assigned to each group (i.e., each row of each Latin Square).

MOS data from sections 3 and 4 was combined in our analysis, although we did distribute the raw data to allow participants to make comparisons within sections, should they wish. It is planned to publically distribute the raw data, our statistical analysis and the synthetic speech itself (all anonymised) via the Blizzard pages on the SynSIG website [2].

Note that for calculation of WER in Section 5, allowance was made for certain spelling variations in listener responses, both because some words were obscure and because many listeners were non-native speakers. Additional problems faced during computation of WER include splitting or compounding words (e.g., "thunder showers" should be considered a correct response, even if the correct transcription is "thundershowers"). Calculation of WER was performed automatically by using a spelling-comparison program written specifically for this purpose. The program was carefully tuned empirically so that the program's decisions on spelling errors were close to the opinions of the experimenters.

As in previous years, system names were anonymised in all the distributed results. Actual listener responses to sections 1,3, 4 and 5 were also distributed together with a lot of extra background information about each anonymised listener. The information was taken from optional responses to a listener feedback questionnaire presented on completion of the evaluation. See Section 7.4 and Tables 13 to 35 for a summary of this information.

7. Discussion

In this section we discuss issues arising from Blizzard 2007.

7.1. Barriers to participation

One registered participant (a commercial organisation) was unable to enter a system this year due to difficulties in getting the contract for the data agreed. This suggests that, in future, all other things being equal, freely available data should be preferred. However, the availability of high-quality data from ATR was a very significant benefit to Blizzard 2007, and we feel this outweighed the legal overheads when using such commercial data.

One group of MSc students registered but did not submit an entry, possibly because of difficulty in raising the USD 500 entry fee. Reduced or zero fees for student entries would solve this problem, but may lead to a larger number of (perhaps lower quality) entries, which may add little to the scientific goals of Blizzard.

7.2. Quality of entries and the aims of Blizzard

Several listeners complained about poor quality synthesis; it is possible that this is a reason for many non-completions of the listening test. Also, such entries give a poor impression of the quality of TTS available today and may lead to a ceiling effect in listener responses, making the better systems harder to differentiate.

Therefore, some thought must be given to the goals of Blizzard, and a balance must be found. Blizzard is

- a scientific enquiry, not a competition
- a comparative survey of current synthesis systems and the techniques they employ
- a valuable opportunity for participants to obtain extensive listening test results for their system, and comparisons with many other systems, which they are unlikely to be able to arrange on their own
- a comparison of techniques, both widely used and novel ones, not just of participants' own engineering skill

We intend to continue to encourage more teams to participate in Blizzard. With increasing numbers, it will be necessary to re-think the listening test design, since the current Latin Square method will probably not scale up beyond about 20 participants.

It may become necessary or desirable to conduct a two-phase listening test. An initial phase would provide the statistics that are currently available for all systems. A secondary phase would take a subset of systems (either the 'best' ones or a representative subset of all systems) forward for a more detailed evaluation.

One goal of Blizzard that is not currently being achieved is to determine *why* some systems are rated higher than others. We have made a first step towards this with the MDS section in Blizzard 2007, but would like to go much further in future.

7.3. Listener recruitment and completion rates

Registration numbers and completion for speech experts increased this year: 163 out of 202 registered speech experts completed the evaluation in 2007 compared to 83/134 in 2006. On the other hand, the rate for volunteers decreased: 65/198 (2007) compared to 113/214 (2006). These comparisons may be misleading however, since some of the 2006 volunteer listener type should perhaps have been registered as speech experts [4]. As noted in Section 6 we used all listener responses to compute the summary statistics in this year's analysis - responses from both complete and partially completed evaluations. A detailed breakdown of the numbers of each listener type whose responses were used in the results for each voice is shown in Tables 5 to 12.

The registration levels and completion rate for US students was very low this year: 16/27 (2007) compared to 44/55 (2006). This is thought to be because the evaluation was based in the UK, and recruitment of US undergraduates was performed with help from US-based faculty, which was not effective. We had hoped to compare responses between the newly-introduced UK undergraduate group with the US undergraduate group in order to determine if there were any systematic differences. The low number of US Undergraduates makes this impossible. However, we still believe that the UK undergraduates are a valid listener group. Conducting paid evaluations in a supervised lab also enabled us to achieve 100% completion rates for the UK undergraduate group, which suggests that we should use this method more in future challenges. It has the additional advantage of more controlled listening conditions (headphones; quiet distraction-free environment)

We have left the web-based evaluation system online. About 20 listeners have registered in the 2 months since the official end of the evaluation and 10 of them have completed it. A new listener group has also been created for current research at CSTR into older listeners' perception of speech synthesis. For the listeners in this group, we have extensive additional information, including highly-detailed audiological test results. We plan to report the results from this listener group at a later date.

7.4. Listener feedback

On completing the evaluation, listeners were given the opportunity to tell us what they thought through an online feedback form. This was based heavily on the Blizzard 2006 listener questionnaire [4], to which we added questions about the new sections we had introduced for Blizzard 2007, about the level of English of non-native listeners, the number of sessions required to complete the evaluation, whether the whole evaluation was taken in the same environment, and the noise level whilst taking it. All responses were optional. Feedback forms were submitted by all but one of the 306 listeners who completed the evaluation (Table 13), and included many detailed comments and suggestions from all listener types.

Listener information and feedback is summarised in Tables 2 to 35. There were more than twice as many male listeners as female (Table 2); the number of native speakers of English and non-natives was almost equal (Table 4). The most frequent first languages (Table 1) of non-natives were Japanese (29), Chinese(21) and German (21).

Most listeners used headphones, (Table 20), most were in the same environment for all samples (Table 21), mostly a quiet environment (Table 22), and most did the evaluation in one session (Table 23). This was good because these are the kind of factors that we cannot control in an online evaluation and the majority of listeners reported using a set-up similar to that which we would have used if we were conducting the evaluation in a lab. Details on the most widely used web browser will be useful when considering configuration issues in the next Blizzard Challenge, (Table 24) though we cannot tell if the browsers used represent listeners' first choice: some comments implied that people had used the browser stated because it worked better with our interface than the one they usually preferred.

Listeners were asked if they found the tasks easy or difficult, and in the latter case to give reasons why. They were also asked about the average number of times they listened to samples in each section (Tables 25 to 35). About 75% of listeners found sections 1 and 2 easy, and about 86% found sections 3 and 4 easy, but about 47% of listeners found section 5 hard. This is reflected by the number of times samples were listened to: about 85%-90% listened to the samples in sections 1-4 just once or twice, but in section 5 nearly 50% listened to the samples 3-5 times and about 15% listened 6 or more times. From comments left about the tasks it was clear that in Sections 1-4 several listeners had doubts about initial calibration of the scale, the size of the scale, and what the instructions meant by 'similar' (section 1) and 'natural' (Sections 2-4). This is a typical problem for listeners doing these kinds of tests. Some suggested that actual examples should have been given to illustrate the scale, but we wanted to avoid imposing our own subjective choices with respect to this, in particular because in Section 2 we wanted to identify the features that listeners themselves appeared to focus on in order to define naturalness. The comments about these issues from all listener types showed that they gave serious thought to the task. Some listeners felt confused by the instructions, although we had expended considerable effort on the wording in order to avoid ambiguity. That the task itself is also unfamiliar for many listeners made this more difficult.

At the end of the feedback questionnaire, listeners were asked to state what they liked most and least, one thing they would change in the evaluation, and for any additional comments. There were many positive comments about the evaluation interface, simple layout, clarity of instructions, use of embedded media players, length and variety of tasks, and being able to stop at any point and do the evaluation in more than one session. Concerning the samples themselves, listeners were impressed by the variety of systems and techniques and how good/convincing/natural the better samples were, but some complained about the inclusion of poor samples which they found made the task more tedious. Several listeners would have liked more feedback of progress within sections. We had intended to include this in response to listeners' comments on previous evaluations and it should be a feature of the next interface.

Section 5 (SUS) was most often singled out as the favourite section by native speakers, who often found the sentences hilarious. For non-natives, it was the most difficult section however, and some suggested that this section should be for native speakers only, due to the obscure vocabulary. Other suggestions for the SUS test included varying the structure of SUS sentences, and having SUS samples from the original speaker so that the WER would also be calculated with natural speech. A WER result for natural speech would of course be extremely useful in interpreting the WERs for synthetic speech.

7.5. Suggestions for future Blizzard Challenges

General listener and participant suggestions for future challenges included excluding systems with really poor quality samples and extending the period of evaluation. Concerning data, there were calls for

- Languages other than English
- A female voice
- Non-US accents

Participants were divided on whether to use larger databases or not. With respect to the content of the evaluation, suggestions included

- Synthesis of paragraphs
- AB comparison tests
- Expressive or emotional speech

8. Conclusions

Three Blizzard Challenges have now been completed and from these we have been able to learn much both about techniques and evaluation methods for speech synthesis. Blizzard is not a competition - it is a challenge with scientific aims. Rather than repeating Blizzard in a similar format next year, perhaps we should now redesign the challenge to investigate difficult areas that have not been included so far, in order to motivate new and interesting approaches.

9. Acknowledgements

Special thanks to the previous organisers of the Blizzard Challenges, in particular Alan Black, Tina Bennett and Brian Langner, for their invaluable advice and software. We would also like to thank Rob Clark, Cassie Mayo, João Cabral, Sebastian Andersson, Dong Wang, and the other staff and students in CSTR for their invaluable assistance in running the Blizzard Challenge 2007. We are very grateful to ATR for supplying the data and to Richard Sproat for generating MRT and SUS sentences. Finally, thanks to all participants and listeners.

10. References

[1] A. Black and K. Tokuda, The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common databases, in Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.

[2] www.synsig.org/index.php/Blizzard_Challenge_2007

[3] C.L. Bennett, Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005, in Proceedings of Interspeech 2005, Lisbon, Portugal, 2005.

[4] C.L. Bennett and A.W. Black, The Blizzard Challenge 2006, Blizzard Challenge Workshop, Interspeech 2006 - ICSLP satellite event, Pittsburgh, Pennsylvania, 2006.

[5] J. Ni, T. Hirai, H. Kawai, T. Toda, K. Tokuda, M Tsuzaki, S. Sakai, R. Maia, and S. Nakamura, ATRECSS - ATR English speech corpus for speech synthesis, Proc. Blizzard Workshop (in Proc. SSW6), August 2007, Bonn, Germany.

[6] J. Kominek, and A.W. Black, The CMU Arctic speech databases, In SSW5-2004, 223-224, Pittsburgh, Pennsylvania, 2004.

[7] www.synsig.org/index.php/Blizzard_Challenge_2007_Rules

[8] A.S. House, C.E. Williams, M.H.L. Hecker, and K.D. Kryter, 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.

[9] C. Benoit and M. Grice, 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.

[10] C. Mayo, R.A.J. Clark, and S. King, Multidimensional scaling of listener responses to synthetic speech. In Proc. Interspeech 2005, Lisbon, Portugal, September 2005.

[11] R.A.J. Clark, M. Podsiadło, M. Fraser, C. Mayo, and S. King, Statistical analysis of the Blizzard Challenge 2007 listening test results. Proc. Blizzard Workshop (in Proc. SSW6), August 2007, Bonn, Germany.

First language	Total
Afrikaans	2
Amharic	1
Bulgarian	1
Catalan	2
Chinese	21
Czech	1
Danish	2
Finnish	3
Flemish	1
French	1
German	21
Greek	2
Hebrew	2
Hindi	1
Hungarian	3
Japanese	29
Korean	8
Persian	1
Polish	5
Portuguese	6
Romanian	1
Russian	2
Spanish	11
Swedish	6
Telugu	1
Thai	2
Turkish	1

Table 1: First language of non-native speakers

Gender	Female	Male
Total	96	205

Table 2: Gender

Age	18-24	25-39	40-59	60 and over
Total	104	155	39	3

Table 3: Age

	Native-speakers	Non-natives
Total	151	149

Table 4: Native and non-native speakers of English

	A	B	C
K	29	20	13
R	53	57	31
S	73	67	40
U	10	6	4
ALL	165	150	88

Table 5: Listener types per voice, showing the number of listeners whose responses were used in the results

Team	Team			General				Technical		
	Blizzard 2006	People	Person-hours	System name	System age	Availability	Hardware platforms	Memory footprint	Sub-word units	Type
Cereproc	yes	4	64	CereVoice	2005	commercial, free academ.	x86, arm powerpc	100 MB	diphone	concatenative
CMU	yes	6	74	Hybrid acoust. unit selection	2006	will be open source	same as Festival	500 MB	diphone	concatenative/ statistical
CSTR	yes	5	8	Festival Multisyn	1996 2002 (u.s.)	BSD-style licence	x86, sparc, powerpc, etc.	603 MB	diphone, phone	concatenative unit selection
DFKI	yes	3	80	Mary	1998	open source	intel, solaris x86, powerpc	300 MB	diphone, half-phone	concatenative
HTS	yes	4	300	HTS	1995	commercial/ open source	windows, linux, mac, PDA, etc.	Voice A: 9-39 MB Voice B: 4-15 MB	context-dependent quin-phone	HMM-based MSD-HSMM
iFlytek	yes	9	600	Interphonic Prototype	2006	prototype to be commer.	PC	1.5 GB	phone	HMM-based unit sel. and wave. concat.
INESC-ID	no	3	40	none	2007	public (CC license)	windows	30-50 MB	phone, syllable	concatenative unit selection
IVO	yes	2	160	IVONA	2003	commercial	win32, unix, mac, pocket pc	2 MB	diphone	concatenative
mXac	no	1	80	CircumReality	2002	public	windows	20 MB	demi-phone	concatenative
NOKIA	no	2	140	NTTS	2005	internal development	linux, nokia tablet 770	no idea	phoneme diphone	concatenative
SVOX	no	3	50	SVOX	1990 2002 (u.s.)	commercial	mobiles and automative	7 MB (Blizzard)	confid.	concatenative
Toshiba	no	conf.	conf.	confidential	confid.	confidential	confidential	confiden.	half-phone	concatenative
UCD	yes	1 + 8 list.	1	Jess	2004	will be open source	FreeBSD, linux	20 MB + voice	diphone	concatenative
UPC	yes			Ogmios	1996	commercial	linux, win pocket pc	500 MB	context-dep. demiphone	concatenative
USTC	yes	6	100	none	2004	no	PC	40 MB	phone	HMM-based parametric synthesis
Voiceware	no	3	120	VoiceText	2000	commercial	windows, linux, solaris, AIX, etc.	voice A: 280 MB voice B/C: 40 MB	phone, half-phone	concatenative

Table 6: Blizzard Challenge 2007 participant questionnaire: part 1.

Team	Voice Building				Components				
	CPU hours	Labelling system	Manual Verif. h/p	Labels	Tools	Lexicon	Prosodic model	Target cost	Join cost
Cereproc	1h	HTK	3h	CMU lexicon	none	own	F0, duration	linguistic (stress, etc.)	acoustic, spectral
CMU	3 days	eHMM	none	phonetic, F0 CMU lexicon	Festival	CMU	Clustergen F0 (unchan. unit)	MCEP	MCEP, F0
CSTR	12h	HTK	none	unilex phoneset	festival, HTK: label., DSP	Unilex	none	12 (different weights)	3 (equal weights)
DFKI	20h	Sphinx	0	CMU lex. (sampa): PoS, ToBI, punctu., word rate	FreeTTS	CMUDict V0.4	F0 and duration	phoneme, durat., stress, pauses F0, syllable break	F0 and 12-MFCC
HTS	20 days	none	none	festival phoneset phone, segment., syllable, word, phrase levels	festival: dumpfeat, HTK	festival: CMU (see paper)	MSD-HSMM	none	none
iFlytek	139h	HTK	400h	pronunciation, prosodic (ToBI)	HTK: segment. HTS: ML train.	CMU	F0 and dur. HMM for u.s.	likelihood of acou. and dur. mod.	likelihood of acou. and concat. mod.
INESC-ID	4-5h	HTK	none	phonetic CMU lexicon	HTK, TCL Snack	CMU	dur. and F0 for target cost	see paper	MFCC
IVO	4 weeks	Sphinx (modified)	none	phonetic: CMU lex. prosodic: F0, power, duration	no	CMU	F0, duration, power	F0, dur., stress, phone position, phonetic context	spectral, F0, power, duration, voiceness
mXac	60h	own	12h utterance transcrip.	phones and words timing, pitch, (own label format)	no	modified CMU lexicon	F0, energy, duration (see paper)	ASR score, F0, ΔF0, energ., duration, context	spectral, contiguos units preferred
NOKIA	10h	HTK	none	phoneme bound. (similar to CMU)	Festival (text analysis)	own	none	context information	pitch, spectrum
SVOX	12h	HMM with Python interf.	1h (30 first utt.)	confidential	no	own	confidential	confidential	confidential
Toshiba	conf.	HTK	confid.	phonet., PoS, ToBI dependency parses	no	own and CMU	dur. & F0: pause, break, accents	F0, dur., phon. context, etc.	spectrum, etc.
UCD	15h	Julian and HTK	0	IPA, PoS, F0, inten., artic. acous. param.	none	Celex	none	basic intonation curve	1-12 MFCC F0, intensity
UPC	5h	Ramses (UPC system)	none	phonetic (sampa) lex. stress, pauses	none	Unilex+ LC-STAR	F0 cont. selec. (target cost)	lexical stress, dur., F0, intens., etc.	spectral, intens., F0, voiceness
USTC	182h	iFlytek	none (iFlytek)	labels provided by iFlytek	HTS: ML train. and param. gener.	iFlytek	HMM-based: F0, duration	none	none
Voiceware	24h	VoiceEZ (HMM ASR)	84h	phonetic, CMU lexicon	none	own	target cost: F0 and duration	phonetic context, F0, duration	spectrum, F0, energy

Table 7: Blizzard Challenge 2007 participant questionnaire: part 2.

Team	Data			Signal Processing				Opinions			Future		
	Extra data for training	Pruning	Voice C	Spectrum	Source	Pitch-marks	Signal modification	Systems' best quality	Strongest comps	Weakest comps	Blizz. 2008	US or other	More data
Cereproc	none	non-Eng.	greedy algorithm	confid.	confid.	confid.	not for blizzard	no compress., own data., pitch smooth.	unit selection	compress. artifacts	yes	yes	no
CMU	CMUDICTs LTS rules	none	none	wavef.	-	EST	none	yes	acoustic unit sel.	joins discont.	yes	yes	no
CSTR	LTS rules	non-Eng., bad dur. phones	greedy text selection	RELP	LPC	EST	OLA	yes	join cost, lexicon, wave. gener.	labelling, DSP	yes	UK	no
DFKI	PoS: WSJ	non-Eng., unk. words	custom method	PCM audio	-	Praat	none	yes	-	automatic labelling	don't know	yes	no
HTS	ARCTIC for MSD-HSMM	none	no	Straight mel-spec.	MBE	no	none	no, more time for training	stoch. mod. (WER)	stochastic model	yes	yes	no
iFlytek	LTS, utt. bound., ToBI	none	greedy algorithm	mel-cepstrum	none	no	none	no, labelling improvement	cost func. from HMM	robust, footprint	yes	yes	yes
INESC-ID	lts, syl. bound. using CMU	none	described in paper	MFCC	-	EST	none	no, starting system	-	-	yes	other	no
IVO	CMU dict. CMU LTS	none	none	original signal	-	own tech.	duration (low scale)	yes	USLTM	labeling	yes	yes	yes
mXac	LTS: lexicon PoS: Project Gutenberg	none	most common units/seq.	custom (see paper)	custom (see paper)	none	F0, dur., spectrum	no (see paper)	ASR to u.s. (robust, automatic)	ASR (muffled formants)	yes	both	no
NOKIA	text phoneme	none	no	MFCC	LPC	EST	none	unclear	unclear (in develop.)	text analysis	yes	both	yes
SVOX	none	see paper	greedy algorithm	confid.	confid.	confid.	confid.	better transc., segm., datab.	confid.	confid.	don't know	both	no
Toshiba	LTS, stress, pronun., etc.	none	random selec. utt.	none	none	yes	see paper	confid.	confid.	confid.	don't know	DE,JP., CN	yes
UCD	Celex (dict.) HTK, TIMIT: feat. extrac., acous. model	no	no	CELP	all-pole filter	none	silence durations normalized	no, better with CMUDICT	synthesizer (articulat. features)	forced alignment dictionary	yes	yes	yes
UPC	LTS: Unisyn lexicon POS: WSJ corp.	10%: ph-dep thres align prob	greedy algorithm	waveform	none	Praat	TD-PSOLA (big devia.)	no, other language	Prosody model	signal processing	yes	other EU	yes
USTC	none	none	iFlytek (corpus design)	LSFs	pulse+ phase modif.	none	Straight Analysis/ Synthesis	no: poor labs for English	MGE train., LSP formant enhancement	param. synt. (muffled speech)	yes	CN	yes
Voiceware	LTS rules (own lexicon)	none	greedy algorithm	none	none	own tech.	none	no, better female voices	pre-selection algorithm	confid.	don't know	yes	yes

Table 8: Blizzard Challenge 2007 participant questionnaire: part 3.

	Registered	No response at all	Partial evaluation	Completed Evaluation
K	62	0	0	62
R	198	57	76	65
S	202	22	17	163
U	27	7	4	16
ALL	489	86	97	306

Table 9: Listener registration and evaluation completion rates

	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17
K	1	3	2	2	2	1	2	3	2	2	1	1	2	1	1	3	0
R	2	3	2	3	3	8	3	3	3	2	2	3	3	1	1	6	5
S	4	3	7	8	2	5	6	4	3	3	4	4	2	4	3	7	4
U	1	1	1	0	2	1	0	0	0	0	2	0	1	0	1	0	0
ALL	8	10	12	13	9	15	11	10	8	7	9	8	8	6	6	16	9

Table 10: Listener groups - Voice A, showing the number of listeners whose responses were used in the results - i.e. those with partial or completed evaluations

	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11	B12	B13	B14	B15	B16	B17
K	2	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1
R	0	3	6	1	4	1	6	4	1	3	4	5	5	1	2	4	7
S	4	3	4	5	5	3	5	3	4	3	2	4	6	6	4	4	2
U	0	0	2	1	0	0	0	1	1	0	0	0	1	0	0	0	0
ALL	6	7	13	8	10	6	13	9	7	7	7	10	13	8	7	9	10

Table 11: Listener groups - Voice B, showing the number of listeners whose responses were used in the results

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12
K	0	1	1	1	2	1	1	1	2	1	1	1
R	0	2	5	5	0	0	4	5	3	2	3	2
S	4	3	5	4	3	3	4	2	2	3	4	3
U	0	1	0	0	0	1	0	1	0	1	0	0
ALL	4	7	11	10	5	5	9	9	7	7	8	6

Table 12: Listener groups - Voice C, showing the number of listeners whose responses were used in the results

Listener Type	K	R	S	U	ALL
Total	62	65	162	16	305

Table 13: Listener type totals for submitted feedback

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate
Total	21	34	78	102	66

Table 14: Highest level of education completed

CS/Engineering person?	Yes	No
Total	194	108

Table 15: Computer science / engineering person

Work in speech technology?	Yes	No
Total	176	124

Table 16: Work in the field of speech technology

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
Total	82	53	34	58	40	9	23

Table 17: How often normally listened to speech synthesis before doing the evaluation

Dialect of English	Australian	Indian	UK	US	Other
Total	4	2	82	47	8

Table 18: Dialect of English of native speakers

Level of English	Elementary	Intermediate	Advanced	Bilingual
Total	18	42	69	20

Table 19: Level of English of non-native speakers

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
Total	241	38	14	4

Table 20: Speaker type used to listen to the speech samples

Same environment?	Yes	No
Total	289	6

Table 21: Same environment for all samples?

Environment	Quiet all the time	Quiet most of the time	Equally quiet and noisy	Noisy most of the time	Noisy all the time
Total	180	90	21	3	1

Table 22: Kind of environment when listening to the speech samples

Number of sessions	1	2-3	4 or more
Total	195	80	21

Table 23: Number of separate listening sessions to complete all the sections

Browser	Firefox	IE	Mozilla	Netscape	Opera	Safari	Other
Total	98	168	5	2	3	12	6

Table 24: Web browser used

Section 1	Easy	Difficult
Total	218	79

Table 25: Listeners' impression of their task in Section 1

Problem	Scale too big, too small, or confusing	Bad speakers, playing files files disturbed others, connection too slow, etc	Other
Total	0	2	44

Table 26: Listeners' problems in Section 1

Number of times	1-2	3-5	6 or more
Total	255	41	1

Table 27: Number of times listened to each example in Section 1

Section 2	Easy	Difficult
Total	220	75

Table 28: Listeners' impression of their task in Section 2

Problem	Unfamiliar task	Instructions not clear	Bad speakers, playing files disturbed others connection too slow, etc	Other
Total	22	25	0	30

Table 29: Listeners' problems in Section 2

Number of times	1-2	3-5	6 or more
Total	269	28	0

Table 30: How many times listened to each example in section 2

Section 3 and 4	Easy	Difficult
Total	253	39

Table 31: Listeners' impression of their task in Sections 3 and 4

Problem	All sounded same and/or too hard to understand	1 to 5 scale too big, too small, or confusing	Bad speakers, playing files disturbed others, connection too slow, etc	Other
Total	1	27	0	19

Table 32: Listeners' problems in Sections 3 and 4

Number of times	1-2	3-5	6 or more
Total	263	33	0

Table 33: How many times listened to each example in sections 3 and 4?

Section 5 (SUS)	Usually understood all the words	Usually understood most of the words	Very hard to understand the words	Typing problems: words too hard to spell, or too fast to type
Total	22	135	121	16

Table 34: Listeners' impressions of the task in Section 5

Number of times	1-2	3-5	6 or more
Total	112	140	42

Table 35: How many times listened to each example in section 5