The Jess Blizzard Challenge 2007 Entry

Peter Cahill, Julie Carson-Berndsen

School of Computer Science and Informatics, University College Dublin, Dublin, Ireland
peter.cahill@ucd.ie, julie.berndsen@ucd.ie

Abstract

The Blizzard Challenge is an independent evaluation of synthetic speech systems. This paper describes the Jess system, one of the participants in the Blizzard Challenge 2007. One of the unique features of the Jess system is the use of articulatory-acoustic feature detection to ensure a higher degree of acoustic consistency at joins. To make results comparable with the system presented last year all components of the system except the actual synthesiser were the same as last year's entry.

1. Introduction

The Blizzard Challenge is an independent evaluation of synthetic speech systems [1]. This year's evaluation used a relatively large (8 hour) speech database. Participants had to use the supplied voice data to build up to three voices:

- Voice A - Used all voice data
- Voice B - Used arctic subset of voice data
- Voice C - (optional) Used participants choice of subset, based on transcriptions and not audio content

400 utterances from 5 genres had to be synthesised using each voice. The 5 genres were: text from stories (*novel*), text from news stories (*news*), conversational speech (*conv*), phonetically confusable sentences (*mrt*) and semantically unpredictable sentences (*sus*). The *conv*, *news*, and *novel* genres each contained 100 utterances and the *mrt* and *sus* genres contained 50 utterances each.

The Jess system is a synthetic speech system designed to be used for synthetic speech research. The system was designed to be straightforward for a non-expert user, without limiting its flexibility for experts to fine tune different aspects to achieve an optimal performance. The synthesiser and its associated suite of tools all use Unicode throughout, including the use of the International Phonetic Alphabet (IPA) where appropriate.

Although synthetic speech systems already exist that can produce high quality synthetic speech, the quality achieved is often proportional to the quantity of manual input required by the system. Much of the work in the Jess system has been aimed at achieving the best possible quality using automatic methods only. Manual input is possible but expertise in synthetic speech should not be a requirement for building voices and synthesising speech.

The first prototype of the system was one of the entries in the Blizzard Challenge 2006 [2]. Participation in 2006 helped highlight the differences between the Jess prototype and other research systems and assisted in identifying where future work should focus. The Jess entry for the Blizzard Challenge 2007 contained numerous improvements on the previous prototype, which are discussed further in Section 3.

Our focus was on voice A, the voice containing all of the voice data. We did not submit the optional voice C as our aim is for the synthesiser to identify which subset of the database is best for synthesis. In doing this the synthesiser uses the phonetic and phonological data that is included in the built voice.

The remainder of this paper discusses the system entered and the results of the evaluation. Section 2 describes the voice building process, Section 3 describes the synthesis system, Section 4 discusses aspects of the evaluation. Results are discussed in Section 5, and Section 6 concludes and describes future work.

2. Voice Building Method

The voice building method is fully automated. A single application is used to perform the complete building of a voice. Building a voice requires a speech database containing speech audio files and a corresponding orthographic transcription of the speech. The user has to specify the location of the voice data as well as the name of the language model to be associated with the voice. It is possible to use different language models for each language.

The voice building application calculates all intermediate voice data during the voice building process and stores it in the speech database. This allows for the intermediate data (*mel frequency cepstral coefficients (MFCCs), fundamental frequency, intensity, $f_1, ..., f_4$, articulatory acoustic features, HTK parameters for forced alignment, a dictionary file for forced alignment and the forced alignment temporal endpoints*) to be modified at a later stage if required. It is possible to update a previously built voice to reflect the modified contents of the intermediate data.

The speech audio is force aligned using the orthographic transcriptions to obtain the words present in each file. A lexical representation is then estimated using the language models and the resulting data is converted into a dictionary for the forced alignment. The acoustic model used in the evaluation was trained on TIMIT [3] using HTK.

The voice building process compresses the speech data using a variation of code excited linear prediction (CELP) encoding implemented through the speex codec [4]. Speech annotations are automatically created by applying the selected language models to the normalised orthographic transcriptions. Each word in the transcription has their appropriate part of speech (POS) calculated using [5] and pronunciation estimated using the selected language model. The current language models are all c4.5 decision trees [6]; however, the system also supports models based on other techniques such as support vector machines (SVMs).

After the voice building application has processed all of the voice data, a single file is produced containing all data needed for synthesis using the voice. The language used for building the voice must be installed with the synthesiser, as the language models are not included in the final voice. This is intentional to ensure the same language models that were used in building the voice will be used for the synthesis to keep the pronunciations consistent.

The built voices were 158MB for voice A (the full voice data) and 18MB for voice B (the arctic subset of the voice data). The voice sizes represent all of the available voice data being included so that different synthesis techniques could be used without any need to rebuild the voice. If only the voice data for the synthesiser component used in the evaluation was included in the voice files their size would decrease by approximately 50%. To make the results of the new synthesis component in the system more comparable with last years entry the same language models and forced alignment acoustic models as last year were used.

3. System Overview

The Jess system is a complete speech synthesis system. It uses an object oriented programming paradigm throughout and was developed in C. Current versions build on many Unix style platforms such as FreeBSD and Linux, using the current source a Windows port is feasible. The system makes extensive use of objects to create a layer of abstraction, resulting in the easy integration of alternative algorithms for algorithm comparison.

The system was designed to support multiple languages and voices as described in Section 2. Each voice is stored in a single file that contains all relevant data for that voice. The voice file was designed to include all speech data that any synthesiser component may need to use. This allows for a direct comparison of synthesis techniques without needing to build separate voices for each synthesis algorithm.

The synthesiser component used in the evaluation was a unit selection speech synthesiser, which creates speech by concatenating real speech segments to form the target speech, in line with [7, 8].

This years entry used a different synthesiser to that of the 2006 Jess entry [2]. To make this years synthesiser comparable to the previous entry, all of the components used other than the synthesiser were kept the same. The language models were trained on the Celex English dictionary [9] and the phone set used was the Celex phone set mapped into the IPA. The motivation for mapping the phones into the IPA is that the system can then contain additional language independent phonological knowledge of the phonemes. As different phone sets are mapped into the IPA, specific details of the phone labels being used do not need to be supplied to the system.

The synthesiser used a Viterbi search to find the optimal path through a sequence of diphones. The optimal path was identified by using the Mahalanobis distance measure as the join cost function with a vector of parameters calculated from the speech data, such that:

$$D(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})} \qquad (1)$$

where $\vec{x}$ is the vector of parameters estimated from the speech of the left part of the join, and $\vec{y}$ is for the right part of the join. The covariance matrix (Σ) was calculated during the voice building process. All math used in calculating the join cost and the covariance matrix was done with 64bit precision. The parameters used in the join cost vectors are 12 MFCCs, fundamental frequency and intensity.

All phones in the voice database have an articulatory-acoustic feature quality associated with them. This was calculated by extracting the articulatory-acoustic features present in the audio waveforms. The articulatory-acoustic features used for the voices built from the voice data were: *voiced*, *vocalic*, *consonantal*, *anterior*, *continuant*, *coronal*, *high*, *back*, *strident*, and *sonorant*. These were chosen from a larger set of feature extractors as they were the most accurate. The articulatory-acoustic feature quality of each phone is used to improve the acoustic consistency present at unit joins. For a more detailed description of the use of articulatory-acoustic feature extraction in speech synthesis see [10].

4. Discussion

Our primary motivation for entering the Blizzard Challenge was to obtain an independent evaluation of the current prototype of our system. Designing a test as large as the Blizzard Challenge is very resource intensive, one of the most important aspects of the evaluation is the access to a large voice database. Public domain databases are typically around one hour in duration and lack the phonetic coverage available when using a database as large as the 8 hour voice data made available in the evaluation. Another factor is that the evaluation has more listeners participating that we would have access to otherwise.

The synthesis algorithm performs synthesis faster than real time. The most computationally expensive stage of the process is the use of the Mahalanobis distance measure, as it involves matrix math for every possible join. No form of clustering [11] is currently implemented as the system performs reasonably fast without it.

The Blizzard Challenge was the first large test performed on the use of articulatory-acoustic feature analysis and the new synthesis component of the Jess system.

Before the Blizzard Challenge, our own experiments on the use of articulatory-acoustic feature extraction showed a clear improvement of synthesis quality. At the time of the evaluation however, we were still experimenting with the optimal configuration of the articulatory-acoustic feature extractors. As the use of the features will result in some units in the voice being avoided during synthesis, configuration of the relevant components primarily depends on the size of voice database. Other factors include the reliability of the results of the feature extractors and it is possible that their quality varies between different speakers and recording conditions. The feature extractor accuracy on test data and the amount of feature extractors being used will also have some influence on their performance.

In situations where the phones with the most common articulatory-acoustic features present is a small set of the phones with that label, a threshold can be set to disable the use of the articulatory-acoustic feature qualities on that set of phones. This is done to avoid over-pruning the available phones. At the time of the evaluation we did not have many experiments done on comparing performance with different thresholds. As voice A was a quite large voice, a threshold was set that would stop over 60% of phones being avoided during synthesis. For voice B, which was only about an hour in duration the threshold was set to be 30%. The motivation for setting the thresholds quite high was because of the large amount of articulatory-acoustic features being used. The comparison of the features present in any phone is explicit, if a smaller set of feature detectors were used much smaller thresholds can also be used as there would be less disagreement when examining the most common features present.

As the aim of the Jess participation was to measure the system performance when performing a completely automatic voice building and synthesis, all utterances were synthesised using a batch synthesis option. The batch synthesis required a list of utterances to be synthesised along with corresponding audio output files.

The aim of our participation was to obtain an independent measurement of the performance of the new synthesis component that is comparable to the Jess system results from the Blizzard Challenge last year. In the near future we intend to perform a test using the same synthesis component as used in this year's Blizzard Challenge, but with improvements to some of the language components, namely:

- Using CMUDICT as a dictionary
- Using UNISYN as a dictionary
- Using different forced alignment acoustic models

An analysis of voices built during both this year's and last year's Blizzard Challenges indicate that these elements are now the most significant source of error in the system.

5. Results

Although it was possible to submit three voices for the evaluation, we opted to only submit voice A and voice B. The system was designed to identify the best segments for synthesis in the voice data, so it was ideal for the system to have as much speech data as possible available.

For each of the two voices, results were made available in categories of listeners that participated in the evaluation. There were 4 different listener categories, each with a single letter representation. The listener categories are shown along with their relevant identification letters in Table 1.

Category ID	Listener Category
K	UK paid undergrads
R	Volunteers
S	Speech experts
U	US paid undergrads

Table 1: *The different listener categories in the evaluation.*

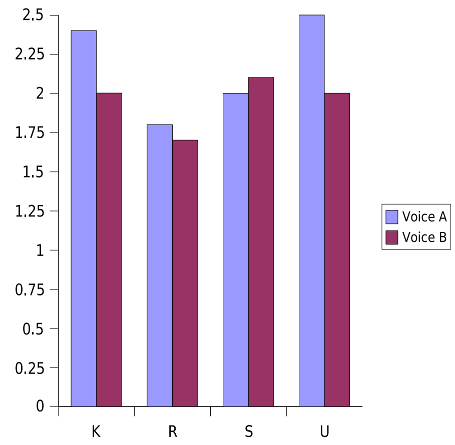

Figure 1: *Mean similarities to original speaker.*

Figure 1 illustrates the similarity of the synthesised speech to the original speaker. The score is on a scale of 1-5, where the actual speaker had an overall mean similarity score of 4.6. Both of the paid undergraduate groups (K and U) have a significantly higher similarity score than the other two groups. The speech expert group (S) was the only group that suggested that the system sounded more similar to the original speaker when the smaller voice was used. The similarity measures are new to the Blizzard Challenge this year so it is not possible to compare them to last year's evaluation.

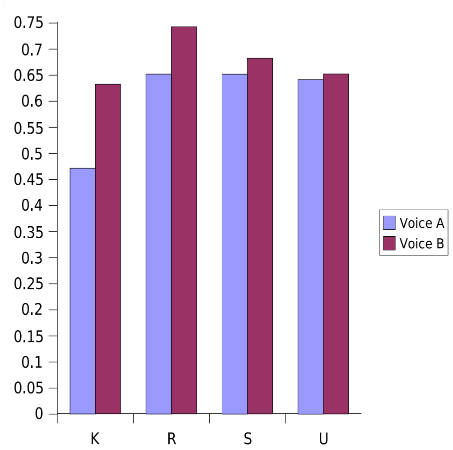

Figure 2: *Word error rate on SUS data.*

Figure 2 shows the word error rate (WER) for each speaker category. WERs for the original speaker are not available. As seen in Figure 2, the results do vary considerably, the UK paid undergrads category being approximately 20% more accurate than the rest. An examination of listeners' input during the WER section of the evaluation shows that errors were often introduced at word endings and at small words that exist as whole words in the voice database. As the small words in the voice database do not require for any segments to be joined to form a word, it indicates that the errors at small words and word endings are due to the temporal endpoints supplied by the forced alignment. The WER results were calculated from the *sus* genre of the evaluation only. This is different than last year, where they were calculated from the *mrt* genre as well. Comparing the WER results of the *sus* genres from this year and last year show an improvement in error rate of over 6%. The WER results of voice A were better than those of voice B, this is a significant difference from the synthesis component used last year.

Figure 3 shows the mean opinion scores (MOS) for each speaker category. The MOS scores are calculated by the participants giving a score between 1-5 of their initial impression of how an utterance sounds. The original speaker got a score of 4.7. The opinions of how each system sounds is always going to be somewhat in comparison to the other systems in the evaluation, so we do not compare the MOS from both years directly. By examining the difference in the MOS between the two voices, voice A had a higher score for all listener categories. Although the MOS scores were under 2.0, a comparison of utterances synthesised by the synthesis components from last year and this year show that this years synthesis component performs significantly better than the one entered previously.

In comparison with the other systems, the overall MOS and median opinion scores show the Jess system came 14^{th} of the 16 participants. This is an improvement on last year's placement and highlights the progress made, as now the system using a fully automatic build of the voice and synthesis of the utter-

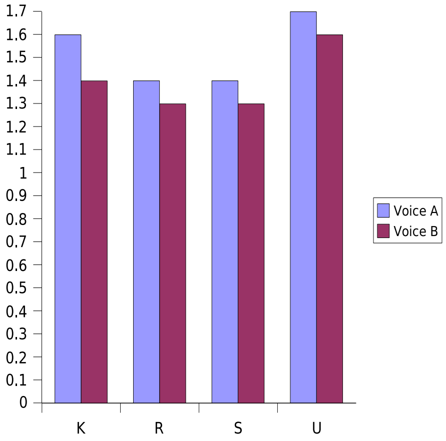

Figure 3: *Mean opinion scores.*

ances came ahead of two other systems.

The results verify that the new synthesis component is performing better than the one used in last year's Blizzard Challenge evaluation. Our focus was on improving the synthesis component and now the weakest point of the system is the dictionary and the acoustic model for forced alignment. Since participation in the Blizzard Challenge 2007 we have started experimentation on using language pronunciation models trained from CMUDICT. While there is a clear improvement when using CMUDICT, we are still fine tuning the models to achieve optimal results.

6. Conclusion

In this paper the Jess entry in the Blizzard Challenge 2007 was described. Our focus was on comparing the current version of the Jess system with the previous synthesis component that we used in the Blizzard Challenge 2006. The submitted entry used an early prototype of our most recent work which incorporates articulatory-acoustic features into the synthesis process.

The entry involved a completely automatic build of the voice and synthesis of the test utterances. No manual tweaking was performed at any stage.

Voices were built from the audio files and orthographic transcriptions only. All other intermediate data required was generated automatically by the system. Two voices were built from the data, voice A consisted of all of the available voice data and voice B consisted of the arctic subset only. Each of the built voices were stored as a single file, voice A was 158MB and voice B was 18MB.

As the majority of work since the Blizzard Challenge 2006 focused on the synthesis component of the system, we decided to use the evaluation as an opportunity to get an comparison of the progress made with the synthesis component. In this year's evaluation all components used were the same as last year with the exception of the synthesis component. A clear improvement is noticed in the results. In comparison with other systems, the current system is performing better than last years entry. There is an improvement of over 6% in the word error rate of the semantically unpredictable sentences.

Participation in the evaluation has helped highlight where our future work should focus, namely on improvements to the language and forced alignment components, as well as to develop the use of the articulatory-acoustic feature extraction further.

7. Acknowledgements

This work was funded by the Irish Research Council for Science, Engineering and Technology (IRCSET) and IBM. The experiments were done using the ATR Blizzard Challenge 2007 speech database. The authors express their thanks to ATR for access to their speech database. The authors also express thanks to Supphanat Kanokphara and Jan Macek for training the feature extractors.

Special thanks go to the Blizzard Challenge organisers for organising the evaluation.

8. References

[1] A. Black and K. Tokuda, "The Blizzard Challenge–2005: Evaluating corpus-based speech synthesis on common datasets," in *Interspeech*, pp. 77–80, 2005.

[2] P. Cahill and J. Carson-Berndsen, "The Jess Blizzard Challenge 2006 Entry," *http://www.festvox.org/ blizzard/ blizzard2006.html*, 2006.

[3] J. Garofolo, N. I. of Standards, and Technology, *TIMIT Acoustic-phonetic Continuous Speech Corpus.* Linguistic Data Consortium, 1993.

[4] G. Herlein, S. Morlat, J. Jean-Marc, R. Hardiman, and P. Kerr, "RTP payload format for the speex codec," *Internet Engineering Task Force*, 2005.

[5] Y. Tsuruoka, "Bidirectional inference with the easiest-first strategy for tagging sequence data," in *Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 467–474, Association for Computational Linguistics Morristown, NJ, USA, 2005.

[6] J. Quinlan, *C4. 5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[7] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, vol. 1, 1996.

[8] A. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Eurospeech*, vol. 95, pp. 581–584, 1995.

[9] R. Baayen, R. Piepenbrock, and H. van Rijn, "The CELEX lexical database (CD-ROM)," *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA*, 1993.

[10] P. Cahill, D. Aioanei, and J. Carson-Berndsen, "Articulatory Acoustic Feature Applications in Speech Synthesis," in *Interspeech*, 2007.

[11] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Eurospeech*, vol. 2, pp. 601–604, 1997.