The Jess Blizzard Challenge 2006 Entry

Peter Cahill and Julie Carson-Berndsen

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

Abstract

This paper describes the version of the Jess system that participated in the Blizzard Challenge 2006. The Jess system consists of a suite of software tools for processing text and speech. The largest component of the system is a multi-platform unit selection speech synthesiser that uses Unicode and the International Phonetic Alphabet (IPA). The system has been designed to be modular so that different synthesiser algorithms can be implemented in a single instance of the system, allowing for alternative techniques to be compared with the same input and target data. In this paper we discuss the algorithms used in the Jess synthesiser to produce speech, insights gained from participation in the Blizzard Challenge 2006, and our intended areas of future work.

1. Introduction

The Jess system is a newly developed synthetic speech system designed for experimenting with different synthetic speech algorithms. A number of different speech synthesis algorithms have been implemented to date, one of which was entered in this years Blizzard Challenge. The Blizzard Challenge is an evaluation that compares the performance of different systems when trained on the same audio databases [1]. The evaluation consisted of two databases, a one hour database and one five hour database. A total of 250 utterances were synthesised using each database. This comprised of 50 utterances per genre, with the genres being: text from stories (novel), text from news stories (news), conversational speech (conv), phonetically confusable sentences (mrt) and semantically unpredictable sentences (sus). The synthesis process used was quite different from conventional unit selection approaches and the Blizzard Challenge 2006 was an excellent opportunity to be able to compare the current version of this approach with other systems.

Corpus based concatenative speech synthesis has conventionally been focused on selecting typically phone sized units of audio from a corpus by minimising the overall cost of the sequence of units [2, 3, 4]. The overall cost is a cost function based on the weighted sum of a target cost function and a join cost function. The Jess synthesiser in this evaluation used a technique that differs from this approach by selecting units of irregular size that have a spectral overlap with the previous and following units. Although the output of the system is a concatenated sequence of units, during synthesis the audio is modeled as sequences of overlapping audio, and all temporal endpoints are decided by the system for each join rather than using phone annotation endpoints. This approach allows for the join function to have a much larger choice of potential areas to join at, and for spectral phonomena to be aligned by the join function.

As this approach is in its relatively early stages of development, we believe that there is much room for improvement. To date much of the work done has focused on making a working prototype of the system. We are currently researching how to improve this technique by using of some of the recent progress in probabilistic approaches to speech synthesis [5, 6, 7].

The remainder of this paper is organised as follows, Section 2 describes the voice building process for the Jess system, Section 3 gives a brief overview of the synthesis concept used, Section 4 discusses the system used in the evaluation, and Section 5 describes the results and insights gained from the perceptual tests and future work we intend to do on the synthesis approach used.

2. Voice Building Method

The system was designed to support multiple voices and languages. Building a voice is an automatic process where the system only requires audio files and plain text files containing an orthographic transcription of what is said in the audio files. The user selects the language of the voice from a list of available languages and then the system will build the voice. Each voice in the system is stored in a single data file that contains audio recordings and annotations, the file format was designed to be useable by different synthesis techniques without requiring that the voice be re-built.

The voice data file contains speech annotations on 4 levels: utterance, word, syllable and phoneme. Annotations are stored in a hierarchical format so that the context of any given unit can be examined. The speech recordings are stored in the form of Mel Frequency Cepstral Coefficients (MFCCs) and Code Excited Linear Prediction (CELP) parameters [8], where the 1^{st} order to the 12^{th} order MFCCs and a variable amount of CELP parameters are used. The final voice data files for the evaluation were 17MB for the one hour database and 85MB for the five hour database.

Speech annotation labels were created automatically from the orthographic transcriptions. Temporal endpoints for the annotations were calculated by performing a hidden Markov model (HMM) based forced alignment on the phone level, using the generated annotation labels. The hidden Markov model toolkit, HTK [9] and Julius [10] were used for the forced alignment process. After the forced alignment, the endpoints for syllables and words were calculated from the phone endpoints. The IPA phoneme labels created are based on the Celex [11] lexical database, assisted by C4.5 [12] decision trees to calculate pronunciations for words that did not exist in the lexical database. Celex had been integrated into the system prior to the evaluation to experiment with English, German and Dutch. However, during the evaluation we realised that we would achieve better results using CMULEX or UNISYN but did not have enough time to add sup-

port for them into the system. The word to phoneme technique that is used at run time is similar to other work such as [13], although the training process used was different.

3. Jess Speech Synthesis System Overview

The Jess system was designed and developed to be a framework for experimentation with new approaches to synthesise speech. The synthesiser component in the Jess Blizzard Challenge 2006 entry is in its infancy when compared to existing unit selection speech synthesis approaches. Unit selection speech synthesis systems to date have typically used either a Viterbi based cost and target function approach or a hidden Markov model based approach. It is common for speech synthesis systems to model the speech as either phones or diphones. The fundamental concept in the synthesiser that differs from the conventional synthetic speech synthesis approaches is that the target speech is modeled as a continuous sequence of audio rather than a sequence of phonological units. By modelling the speech as continuous audio, joins are possible at 10ms windows, where the use of phonological units would result in joins only being possible at unit boundaries. This approach therefore uses a join function that encourages joins at areas of spectral similarity over annotation similarity, so that join points are customised for each join. This approach does also help deal with inaccurate phone endpoints.

The synthesis process is initiated similar to the conventional approach of a target utterance structure being predicted by automated techniques and then suitable unit candidates from the inventory are proposed to the synthesiser component. The estimated target utterance structure consists of word, syllable, and phone labels. Duration modelling was used to predict the ideal phone durations, where the modelling strategy used would select a matching duration to a phone with a similar context in the audio database. This allows for the phones selected to have a similar duration to the target without the need for modifying the speech signal.

From the available sequences of phonological units, the temporal endpoints of each sequence are selected. For the remainder of the synthesis process only the endpoints of the sequence of units and the spectral data contained between the points is used. The synthesiser component will calculate what sequences of continuous units are available for any given point in the target utterance, and attempt to organise these sequences so that the end of one sequence will overlap with the start of the following sequence. If only short sequences are available it is likely that more than one sequence will be available for use in the same location of the target utterance, where as if a sequence of a few words in length is available it is quite likely that it will be the only sequence of that length that is available. The potential sequences are then searched, looking for the most suitable one. The most suitable one being the one with the best join, where a join cost function is used to examine all possible join points and determine which one is best.

The initial stage of the search decides whether or not an overlap exists by examining some spectral properties of the audio. This is done to produce a smaller list of candidate sequences for the join. The fine-grained spectral properties of the smaller list of candidate units is then compared in detail, in an attempt to find the best spectral alignment available.

Figure 1 illustrates how this concept works. Two segments need to be joined at an optimal point anywhere between the two dotted lines, which mark the bounds of the expected overlapping area. A closer examination of the audio between the dotted lines

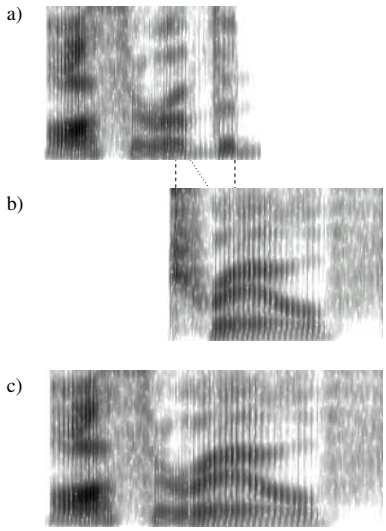

Figure 1: *Sample join. a) is a sequence of speech that needs to be joined to b). c) is the result showing formant continuity during this transition.*

shows that the end of segment ‘a’ is an area of voiced speech with the first four formants clearly visable. The audio segment ‘b’ starts with some frication, but then this changes to voiced speech with the lower formants present in similar locations to ‘a’. The system will calculate that a high quality join is possible here going from the start of the overlapping area in ‘a’ to a point just over half way into the overlapping area in ‘b’. While such a high quality join will not always be available, this approach does seem to give good results. In the conventional Viterbi based approach, a spectral distance comparison at phone or diphone endpoints would result in the join getting a very bad join score as it would be comparing a voiced segment of speech with a fricated area.

Possible overlap areas can be of up to a few words in size, although 2-4 phonemes was the most common with the voices used. The system decides where the ideal join point is, if a stop is detected it will have the highest possible influence on a join, otherwise pitch and MFCC spectral distance are the most influential factors. Pitch is allowed to vary within a range of 20dB and still be considered acceptable. Spectral distance between the MFCCs was measured using the Euclidean metric. After the temporal points have been calculated, the resulting speech will have a natural smoothness due to the spectral measures used during the join function. The use of spectral smoothing techniques [14] at this stage was omitted as it removes some of the natural smoothness that is currently being achieved by this technique.

In the situation where there are two sequences for use, with one following the other (i.e. without any overlapping segments), a diphone unit will be used to give a suitable transition between the two units. In such cases all suitable diphones will be examined with the join function to find the best match, where the first half of the diphone will be overlapping with the end of the first sequence and the second half will overlap with the start of the following sequence.

The system uses CELP parameters at the decided join points to reconstruct the original audio and perform a concatenation to produce the final speech signal.

4. Discussion

The Blizzard Challenge was the largest test that the Jess system has participated in to date. Development of a balanced large scale test is very time consuming and the Blizzard Challenge was of great benefit to the Jess system in this respect. Participation alone highlighted certain algorithms in need of improvement and also resulted in some implementation improvements. As all testing to date had been performed on the much smaller one hour CMU ARCTIC databases, the use of the five hour database highlighted areas where further algorithm optimisation was required and where increased variable bounds was required. The five hour database also contained longer utterances than had been dealt with previously and the system had to be modified to allow for utterances in excess of 18 seconds.

The process of generating the required data files, such as language and voice (including annotations) was fully automatic. The system generated pronunciation rules from a given lexicon without any user interaction or seeding. The lexicon used for this evaluation was the English phonology wordforms database of Celex [11], where the pronunciation rules would be used if a word being synthesised did not exist in the lexicon.

An advantage of using the approach described was that with the algorithms in use, the amount of processing required increases linearly as opposed to a Viterbi based search where the processing would increase exponentially. We considered using a clustering technique similar to work such as [15], which would also help limit the amount of extra processing time required for larger corpora, but with the given time constraints such an approach could not have been tested properly and it would have been possible that the approach would have resulted in omitting suitable units from the search. We intend to implement a cluster-style technique in the near future.

The linear search for units used allows for the synthesis process to be analysed in a more readable form than that of Viterbi or similar searches where large cost matrices are used. This has helped highlight what parts of the algorithms used sometimes result in sub-optimal units being selected.

Examining units during the search was the most computationally expensive part of the synthesis. In particular, the initialisation of the search, which examines the units to see which ones are continuous had a larger effect on the speed of the synthesiser than all other stages combined. We are currently working on optimising this stage of the process.

5. Results

There were two voices used in the evaluation, a one hour database (ARCTIC) and a five hour database (FULL). The results contained the performance of the system using each of these voices separately. The evaluation was done by three different categories of listeners, random/volunteers (R), speech experts (S) and undergraduates (U). Figure 2 illustrates the difference in mean opinion scores (MOS) over the different listener categories. The MOS scores are on a scale of 1 to 5, where real human speech was given a score of approx 4.5/5 (it varied slightly between each listener category).

It can be seen from Figure 2 that the system has a higher MOS with the FULL database. However, considering that the FULL database is approximately five times larger than the ARCTIC one, we expect that it would be possible to get much better results with the FULL database if the system was tuned for dealing with databases of that size. As testing of the system until now has been

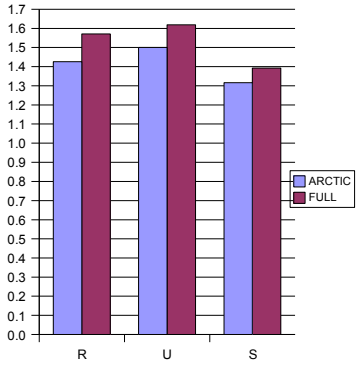

Figure 2: *Overall MOS scores.*

on databases similar to the ARCTIC one, databases of this size seem to be optimal.

The lexical database in use will significantly affect the results. As the target phoneme sequence will be decided by the lexicon and pronunciation rules (which have been trained off the lexicon), any inaccuracies at this level are sure to result in a lower quality synthesis. We intend to perform further tests using the CMULEX lexicon and the UNISYN lexicon, comparing them to the Celex based results that we currently have. To date, tests done indicate that CMULEX does have a more consistent use of vowel phonemes than Celex.

An analysis of the word error rates for each listener category indicates that the synthesis technique used is optimal with smaller databases. Larger databases result in a larger search space, where the longest sequence with the smallest MFCC distance will sometimes be so due to an error in labelling. In such cases a manual analysis of available units suggests that if a shorter, more common sequence was selected the output would be significantly better. In addition to this, it is clear that a form of spectral analysis by the labelling part of the system would also help reduce mis-labelled units from being used. While most labelling errors will be due to the lexicon in use, the actual orthographic transcriptions can also be incorrect at times where a rhyming word may actually have

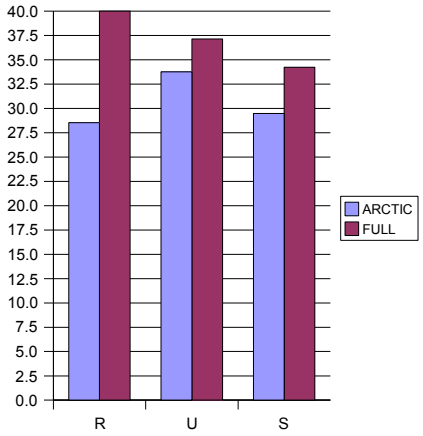

Figure 3: *WER rates for each listener category on mrt data.*

been said by the speaker instead of the prompted word. We expect that spectral analysis would also help solve this problem.

Figure 3 illustrates the word error rates on the mrt data. It shows that the system performed better with the smaller database. Each listener group could understand more words using the ARCTIC database than with the FULL database. This suggests that although the speech may sound more human with the FULL database as seen in Figure 2, the listener actually understands more of the speech produced with the smaller database. This indicates that the current searching algorithm is not using the database optimally. We intend to add spectral parameters into the initial stage of the search to attempt to solve this problem as it would appear that searching for phonological annotations alone is not sufficient. The results suggest that it may be possible that the system tested may perform better with a corpus smaller than the ARCTIC one.

6. Conclusion

Participation the the Blizzard Challenge 2006 was of great benefit to the Jess system. The test utterances highlighted how the system performs when dealing with different situations, and provided us with much insight as to which areas need to be focused on to improve the system most effectively. All previous testing had been done on much smaller databases and as a result the use of the five hour FULL database was an excellent test for both the implementation and the algorithms in place.

The results did show that more listeners understood what was being pronounced when the smaller ARCTIC database was in use. This suggests that our current searching technique was not making optimal use of the speech database.

The use of our alternative synthesis technique has proven itself as a suitable synthesis technique for human sounding speech. Although our current version cannot produce speech at a quality as consistent as some other approaches, we intend to address this problem in future work. The results indicate that the system performs better with smaller databases, so we intend to develop the current algorithms further so that they can use larger speech databases more efficiently.

Improvements and further development has been on going since participation in the Blizzard Challenge 2006. Participation and the results have given significant insight as to what components need to be focused on in future work. We intend to perform tests using the CMULEX and UNISYN lexica; manual inspection of these suggests that better results can be achieved by using either of them. We intend to improve the initial stage of the search as in cases where the resulting speech was difficult to understand, it was often due to the unit candidates calculated by the initial stage of the search.

Another avenue of research is to investigate how probabilistic speech synthesis progress may improve the system, as well as experiment with machine learning techniques to calculate a more detailed target utterance structure.

7. Acknowledgments

The authors would like to thank the Irish Research Council for Science, Engineering and Technology (IRCSET) and IBM for funding this work. Special thanks go to Tina Bennett, Alan Black (and anyone else at CMU) and ATR-SLC for making this evaluation happen.

8. References

[1] Black, A. and Tokuda, K., "The Blizzard Challenge 2005: Evaluating Corpus-Based Speech Synthesis on Common Datasets", Proc. Eurospeech 2005, Lisbon, 2005.

[2] Vepa, J. and King, S., "Join Cost for Unit Selection Speech Synthesis", "Text to Speech Synthesis", pp.35-59, ISBN: 0-13145661-X, 2004.

[3] Hunt, A. and Black, A., "Unit selection in a concatenative speech synthesis system using a large speech database.", Proc. ICASSP, pp. 373-376, 1996.

[4] Clark, R., Richmond, K. and King, S., "Festival 2 – build your own general purpose unit selection speech synthesiser.", Proceedings of 5th ISCA Speech Synthesis Workshop, pp. 173-178, Pittsburgh, PA, USA, 2004.

[5] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, vol.3, pp.1315-1318, June 2000.

[6] Tokuda, K., Zen, H. and Black, A., "An HMM-based speech synthesis system applied to English", 2002 IEEE Speech Synthesis Workshop, Santa Monica, California, Sep. 11-13, 2002.

[7] Zen, H., Toda, T., "An overview of Nitech HMM-Based speech synthesis system for Blizzard Challenge 2005", Proc. Eurospeech 2005, Lisbon, 2005.

[8] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 10, pp. 937-940, 1985.

[9] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., "The HTKBook for HTK version 3.2", Cambridge University Engineering Department, 2002.

[10] Lee, A., Kawahara, T. and Shikano, K., "Julius — an open source real-time large vocabulary recognition engine", Proc. Eurospeech, pp. 1691–1694, 2001.

[11] Baayen, R. H., Piepenbrock, R. and Gulikers, L., "The CELEX lexical database (CD-ROM)", Linguistic Data Consortium, University of Pennsylvania, 1995.

[12] Quinlan, J., "C4.5 : Programs for Machine Learning", San Mateo: Morgan Kaufmann, ISBN: 1558602380, 1992.

[13] Black, A., Lenzo, K. and Pagel, V., "Issues in Building General Letter to Sound Rules", ESCA Synthesis Workshop, Australia, 1998.

[14] Chappel, D. T. and Hansen, J., "A comparison of spectral smoothing methods for segment concatenation based speech synthesis ", Speech Communication, vol. 36, pp. 343-374, 2002.

[15] Black, A., Taylor, P., "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", Proc. Eurospeech 1997, pp. 601-604, 1997.