Multisyn Voice for the Blizzard Challenge 2006

Robert Clark, Korin Richmond, Volker Strom and Simon King

CSTR, The University of Edinburgh, Edinburgh, UK.
(robert|korin|vstrom|simonk)@cstr.ed.ac.uk

Abstract

This paper describes the process of building unit selection voices for the Festival *Multisyn* engine using the ATR dataset provided for the Blizzard Challenge 2006. We begin by discussing recent improvements that we have made to the Multisyn voice building process, prompted by our participation in the Blizzard Challenge 2006. We then go on to discuss our interpretation of the results observed. Finally, we conclude with some comments and suggestions for the formulation of future Blizzard Challenges.

1. Introduction

The *Blizzard Challenge* is a speech synthesis evaluation open to anyone with a speech synthesis system who wishes to participate. This paper describes CSTR's entry in the Blizzard Challenge 2006.

Each entrant to the competition is required to build a voice for their speech synthesis system using the speech data supplied by the organisers. A test set of previously unseen sentences is then released, which entrants are required to synthesise and submit the resulting speech waveforms for perceptual evaluation. A subset of this test set was evaluated alongside contributions from other research groups in a large perceptual experiment.

2. Preparation

2.1. The data set

The data set provided was recorded at ATR (Advanced Telecommunications Research Institute International, Japan) and comprised 4273 utterances recorded by a male speaker with an American accent. The dataset was composed of data from three different genres: a CMU ARCTIC [1] dataset (1082 utterances[1]), a conversations dataset (2134 utterance) and newspaper text (1057 utterances). The data was supplied as mono RIFF encoded waveform files, with 16kHz sample rate and 16 bit precision. In addition, the supplied data included files containing the text for each utterance, a list of phones for each utterance, and a list of aligned phones for each utterance. As we use our own phone set, we disregarded the phone lists and alignments provided, and instead used the wavefiles and recording script text to perform our own forced alignment labelling (described in more detail below).

[1]this is not a complete ARCTIC set as some utterances had been removed to form part of the test set

2.2. Preparing the data

Our aim is to provide as close as possible to an automated voice building solution, and so we tried to avoid manual intervention in data processing as far as possible.

The first step we took in building the *Multisyn* [2] voice was to briefly examine the speech and text data provided, to spot any likely text normalisation problems, gross reading errors or issues of consistency. For example, we decided it was necessary to discard three of the utterances from the data set at this initial stage; two utterances were discarded because the text spoken was entirely Spanish, and one contained a long German name that the speaker had trouble with.

Next, we used an automatic script to reduced the variation in the amount of silence that each utterance contained. Silences of more than 50ms in length, both at the ends of an utterance and internally were reduced to 50ms. We did this because previous experience has shown this provides more robust silence models, and avoids potential problems in certain cases in training, particularly with transition matrix probabilities not summing to one.

We also automatically analysed the text for words that were not in our Unisyn[3] lexicon. We found 746 words we did not have pronunciations for. The majority of these words turned out to be Japanese names, which we decided to not add to our lexicon due to time constraints. The remaining 50 or so words we decided to add to our lexicon, in order to avoid relying on letter-to-sound rules.

3. Segmenting the data

The data was automatically segmented with a forced alignment procedure using HTK [4]. MFCCs were generated using a 10ms window with a 2ms shift; we used 12 MFCC coefficients and log energy, without the standard utterance-based energy normalisation which HTK carries out by default.

3.1. Pronunciation modelling

Prior to the Blizzard Challenge 2006, the *Multisyn* forced alignment procedure used a single phone sequence, output by the Festival front-end processing modules, to align against the speech waveforms. To alleviate the re-

strictions this imposes, single phone substitutions could be defined by the user in order to allow instances of processes such as vowel reduction to be accurately labelled. Cases where these substitutions overgeneralised (i.e., where the HMM alignment wrongly featured a reduced vowel where a reduced vowel was not allowed) were filtered out at the stage of processing the alignment output to build the voice.

For our entry to the current Blizzard Challenge, we have extended the labelling procedure to include all pronunciation variants available in the Unisyn lexicon we use. So, for example, if the word "economic" were to appear in the recording script, the HMM-based forced aligner would be provided with two pronunciation variants to align with the speech waveform: one with an initial [i] vowel, and one with an initial [ɛ] vowel (IPA notation). This method obviously also allows vowel reduction to be accounted for as a sub case. For example, three variant pronunciations for the word "the" are included as options for the forced alignment: [ðə], unstressed [ði] and stressed [ði]. Therefore, the substitutions used previously are no longer necessary, and instead a much more fine-grained specification of pronunciation variation is possible.

Informal evaluations while developing the voice for the Blizzard Challenge 2006 indicated this made a significant difference to labelling accuracy, which translated to a significant improvement in the resulting synthetic voice quality. We intend to measure this quantitatively and report on it in the near future.

3.2. Alignment procedure

Single mixture monophone models with three emitting states were initially trained from a flat start using a fixed most likely phone sequence as a starting point. After three iterations of training, a first alignment, again with the fixed phone sequence, but allowing vowel reduction was performed. This alignment was then used for three more subsequent re-estimation iterations.

The resulting models where then used to produce an alignment from the phone lattices and at this stage a 'tee' model was added to model potential short pauses between words. The models where then re-estimated again, before a second lattice alignment was obtained. Then the number of mixtures was increased to eight, in steps with re-estimation and a final alignment performed from the lattice. This segmentation was used to label the dataset.

We experimented with changing the window size and shift used in generating the MFCC, part in response to suggestions that a 10ms was really too short to capture the frequency resolution that was needed. We tried combinations of a window length of 10 and 15ms with a window shift of 2 or 5ms. Our result suggest that there was little difference between using a 10ms or a 15ms window, but using a 5ms shift provided a labelling where the label times of each phone type were more constant (3.9% outliers rather than 5.0% - 5.5%) but the resulting segmentation resulted in a voice of worse quality, suggesting that the distributions of duration outliers alone is not a good indication of the quality of a segmentation.

4. Building the voice

In addition to the labelling, pitch mark files were generated using *pda*, the pitch tracking program which is part of the Edinburgh Speech Tools, and F_0 files were generated using the ESPS programme *get_f0*. Pitch synchronous LPC coefficients and residuals were then generated to be used at synthesis time.

We built two voices for our Multisyn unit selection engine, one from the full data set and one just from the ARCTIC data. We were reluctant to submit a voice built from just arctic data, as we know this is not a sufficient amount of data to build a good Multisyn voice from. However, we decided that having the evaluation results available for such a voice for comparison to other systems would make this voice a useful contribution.

4.1. Join and target cost specifics

The join cost consists of three equally weighted components for pitch, energy and spectral mismatches. Spectral discontinuity is estimated by calculating the Euclidean distance between two vectors of 12 MFCCs from either side of a potential join point. The MFCC coefficients used for the join cost calculation are taken from the parameterisation produced by the HTK tool *HCopy*, performed during forced alignment.

The target cost comprises weighted components lexical stress, phrase-finality (sentence-internal phrase boundaries are distinguished sentence-final ones), part of speech (content or function word), and position of the diphone in its syllable and word, left and right phonetic context. A penalty is added if either half of the candidate diphone should be voiced (stops and fricatives excluded) but the pitch tracker decided for unvoiced.

Note that apart from phrase boundaries, prosody is not modelled at all, i.e. there are no components for accentuation, F_0 or phone duration.

5. Summary of Evaluation Procedure

Evaluation was carried out using three groups of individuals: speech experts, undergraduate students, and "random people". Each group performed five evaluation tasks. The first three tasks were Mean Opinion Score (MOS) tasks where subjects were asked to rate utterances on a 1-5 scale. The tasks evaluated speech from the three domains included in the database: Newspaper text, conversation, and out of copyright literature (CMU ARCTIC). Natural speech was included in the MOS tasks to provide a top-line for the evaluation. The other tasks were Word Error Rate (WER) tasks, where subjects were asked to type in what they heard. One task was an MRT (Modified Rhyme Task) [5] where subjects were presented with varying single syllable words in a carrier sentence,

and the other task was to transcribe semantically unpredictable sentences [6]. More details of the evaluation procedures can be found in [7].

6. Evaluation results

6.1. WER tests

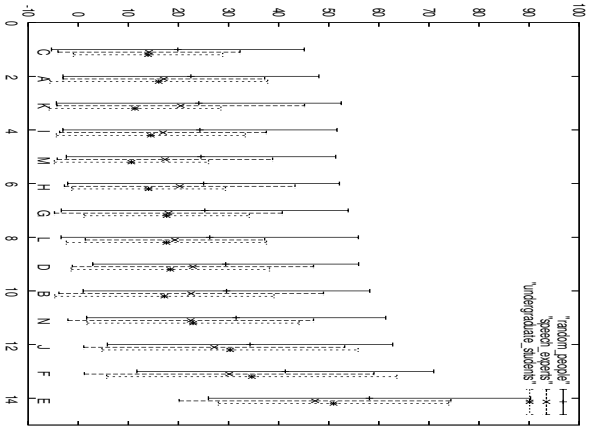

Figure 1: *Mean and standard deviation of word error rates of the systems made of the* **full** *speech corpus.*

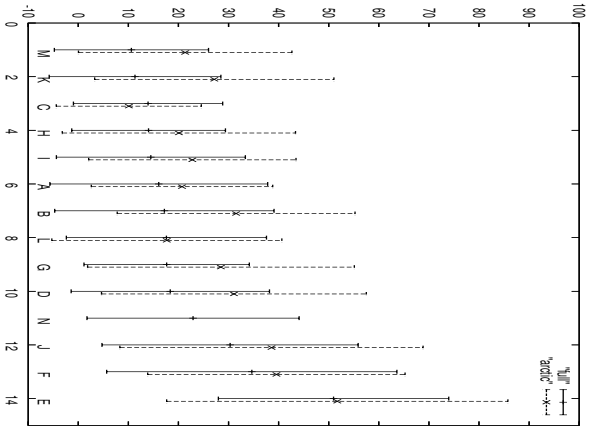

Figure 2: *Mean and standard deviation of word error rates of the systems made of the* **arctic** *speech corpus.*

As means and standard deviations were provided from the WER tests, we decided to assume the data was normally distributed and ran a series of t-tests to compare our own performance with the best performing system in each task. We found that for our voice built from the full data set there was never a significant difference in WER between our system and the best system. For the voice built just from the ARCTIC data there were some instances where our system was significantly worse than the best system for a particular task.

We conclude from these results that with the exception of a couple of systems, there is little difference between the comprehensibility of the different systems.

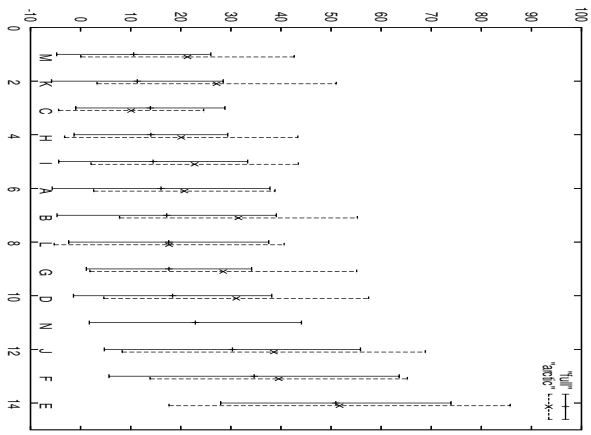

Figure 3: *For each system, compare the undergraduate's word error rate of the version made from the full speech corpus to the version made from the arctic corpus.*

6.2. MOS tests

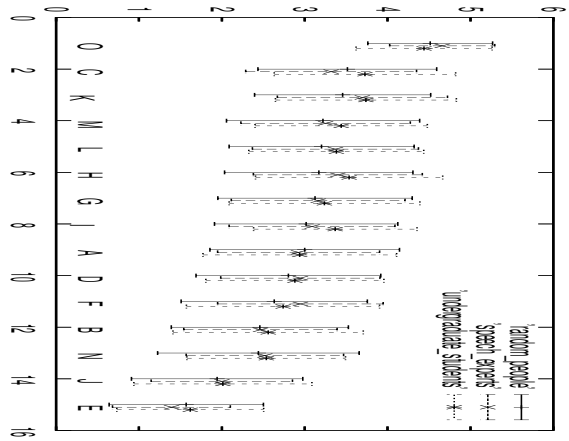

Figure 4: *Mean and standard deviation of MOS scores, given to the systems made of the* **full** *speech corpus.*

For the MOS tests, natural speech was introduced into the evaluate, shown as system **O** in figure 5. Scores for other systems are in about the same range. When comparing the two versions of each system, it looks as if most systems take very little or no advantage from a larger speech data base in terms of MOS, although in terms of WER, most systems improve significantly, by -20% on average, ranging from -2% to -36%.

7. Conclusions

The overall WER results suggest that it is reasonably easy to build a comprehensible speech synthesis system, whereas the MOS tests suggest that building a natural sounding system is a lot harder, and that current systems are clearly perceived less well that natural speech.

Statistical analysis suggests they there may be few significant differences between the WER scores of different systems, but even so there is certainly a reasonable

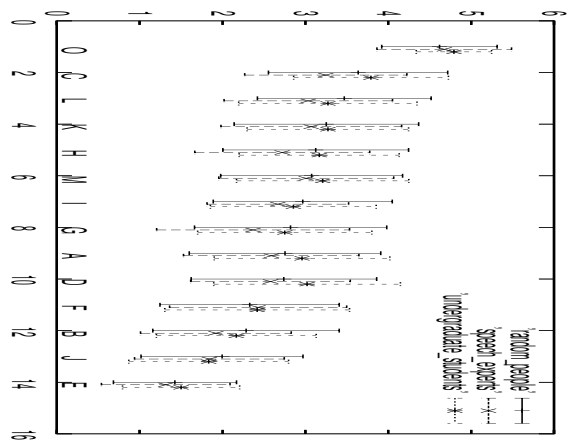

Figure 5: *Mean and standard deviation of MOS scores, given to the systems made of the* **arctic** *speech corpus.*

level of consistency in the scores of each system across the range of tests and subjects.

7.1. Observations from building the multisyn voices

Standing out as one of the most important aspects of unit selection speech synthesis is the need for accurate phonetic segmentation of the voice corpus. In some respects this is counter intuitive, because if joins are being made at diphone boundaries, then the fine-grained accuracy of phone boundaries should not be so important.

The version of Multisyn used here is still without any prosodic modelling as we didn't have time to incorporate recent developments into the system. We believe that caused potential problems with the resulting prosody in some instances, which may have adversely effected our MOS scores in some cases. This however does not seem to affected the comprehensibility of individual words, as shown by the WER results.

We strongly support the notion of continuing to run Blizzard Challenge evaluations on an annual basis. We have found they are useful not only because they provide the opportunity for comparison between different synthesis systems, but also purely because it provides a focus to improve overall synthesis quality and stimulates new ideas. This is probably particularly true for the research oriented entries.

For future Blizzard Challenges, we suggest a variety of interesting directions might be taken. One suggestion would be to attempt a more fine-grained comparison of the systems entered. So far, systems have been compared as a whole; speech data has been provided for groups to process in whatever way they choose. It might be interesting though to compare the subparts of system entries. For example, system entries could be required to all use the same labelling instead of being allowed the freedom to perform their own. In this way, it would be possible to factor out the influence of labelling accuracy between systems and compare them solely on the remaining differences.

We would also like to see other evaluation techniques considered, for example, perhaps forced choice tests in some instances, to improve the likelihood of obtaining significant differences between systems.

Finally, it may prove fruitful to aim future Blizzard Challenges at some of the major questions in speech synthesis other than quality and intelligibility; for example affective speech synthesis, voice transformation, and so on.

8. Acknowledgements

Simon King is supported by an EPSRC Advanced Research Fellowship GR/T04649/01, and Volker Strom is supported by Scottish Enterprise through the Edinburgh Stanford link programme.

9. References

[1] J. Kominek and A. Black, "The CMU ARCTIC speech databases," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224.

[2] Robert A.J. Clark, Korin Richmond, and Simon King, "Festival 2 – build your own general purpose unit selection speech synthesiser.," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 173–178.

[3] Susan Fitt and Stephen Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech '99*, Budapest, 1999, vol. 2, pp. 823–826.

[4] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 3.2)*, Cambridge University Engineering Department, 2002.

[5] A.S. House, C.E. Williams, and K.D. Hecker, M.H.L.and Kryter, "Articulation-testing methods: Consonantal differentiation with a closed response set," in *J. Acoustic. Soc. Amer.*, 1965, vol. 37, pp. 158–166.

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

[7] Alan W. Black and Keiichi Tokuda, "The Blizzard Callenge - 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc Interspeech 2005*, Lisbon, 2006.