Multisyn Voices from ARCTIC Data for the Blizzard Challenge

Robert Clark, Korin Richmond and Simon King†*

CSTR, The University of Edinburgh, Edinburgh, UK.

(robert|korin|simonk)@cstr.ed.ac.uk

Abstract

This paper describes the process of building unit selection voices for the Festival *multisyn* engine using four ARCTIC datasets, as part of the Blizzard evaluation challenge. The build procedure is almost entirely automatic, with very little need for human intervention. We discuss the difference in the evaluation results for each voice and evaluate the suitability of the ARCTIC datasets for building this type of voice.

1. Introduction

This paper describes the process of building and evaluating the four ARCTIC [1] voices for the Festival [2] *multisyn* [3] engine as part of the Blizzard challenge. The build process was almost entirely automatic, and used our standard procedure without modification for this challenge. The following sections describe in detail the build process and comment on the results of the perceptual testing and the usability of the dataset for this type of speech synthesis.

2. Building the voices

The build procedure for each of the four voices was, for the most part, identical. We mention specific voices by name only to highlight differences between the procedures for each.

The four datasets (`bdl`, `rms`, `slt` and `clb`) were downloaded and unpacked. We had previously built a *multisyn* voice using the `bdl` data, but decided to rebuild it from scratch, as changes had been made to the build procedure since it was initially built. We had also previously built a voice from the `awb` data (not part of this challenge), so we had a reasonable idea of what to expect from these datasets. The speech waveform files, pitch mark files and *festvox*-style utterance list were copied for use in the voice building process; all other supplied data was discarded.

2.1. Segmenting the data

Each dataset was automatically segmented by a forced alignment procedure using the *HTK* hidden Markov model toolkit [4].

*Supported by EPSRC grant GR/R94688/01.

† Supported by EPSRC Advanced Research Fellowship GR/T04649/01.

An initial phone sequence was generated by Festival for each utterance in the database, using the same configuration for the language processing component of Festival that was to be used for the finished voice. We chose to use our own American English lexicon, *unilex* (based upon our Unisyn accent-independent keyword lexicon [5]), for each voice. For comparison, we additionally built a second voice from the `rms` dataset using the CMU lexicon. We will call the CMU version RMS and the unilex version we will call RMS2.

The supplied data contained around 50 words, mostly proper names, which were not in our unilex lexicon. Pronunciations for these words were added manually to the lexicon before the initial transcription for alignment was generated.

In addition to the phones for the utterance produced from Festival's front end, this initial transcription also contains initial and final silences, extra labels to represent the closure portion of stops and affricates and short pause labels after each word.

The alignment was carried out using standard left-to-right monophone HMMs with three emitting states and observation densities with eight Gaussian mixture components. The short pause model is a "tee" model: it has a skip transition, so can have a duration of 0 frames. The speech was parameterised using `HCopy` from HTK as 12 Mel-scale cepstral coefficients, energy, deltas and delta deltas. A relatively short window size of 10ms was used with a short 2ms shift in order to produce more precise times for the label boundaries. These settings have been found in previous experiments.

1. models trained from a flat start using the initial transcription and 5 iterations of embedded training
2. an intermediate forced alignment is carried out, in which vowel reduction is allowed
3. models further trained using this intermediate transcription
4. a second forced alignment is carried out with respect to the initial transcription, again allowing vowel reduction
5. models further trained using the second intermediate transcription; observation densities gradually increased to a mixture of eight Gaussian components using HTK's standard "mixing up" procedure
6. a final forced alignment, again with respect to the initial transcription,and allowing vowel reductions,

then produces the labels and boundary times used to build the voice.

2.2. Using the segmentation

The segmentation times from the alignment process were taken and added to the linguistic structure for each utterance. The process involved automatically inserting and deleting silences as dictated by the alignment process, accommodating the labels for closure portions of stops an fricatives (the end of the closure was used to mark the join point for diphones resulting from these segments) and marking reduced vowels where appropriate. Phones without pitch marks were also identified at this stage, as these were not used in the final inventory.

2.3. Finishing the voice

A pitch tracker (CSTR's own *pda* from the Edinburgh speech tools) was used to get F_0 and this was incorporated with the MFCCs and normalised to be used by the join cost. LPC coefficients were also generated to be used as the final voice representation.

3. The synthesiser specification

The *multisyn* [3] implementation in Festival uses a conventional unit selection algorithm. A target utterance structure is predicted for the input text, and suitable diphone candidates from the inventory are proposed for each target diphone. The best candidate sequence is found that minimises target and join costs.

If no examples of a particular diphone can be found in the database a list of suggested alternatives is provided as a backing off solution. These typically look for schwa as an alternative to full vowels, and use silence as a last resort.

The target cost is formulated as a weighted sum of a number of normalised components, which each score how well a candidate matches the given target. These features include (most highly weighted first): lexical stress, position of diphone in the phrase, part of speech (content or function word), position of the diphone in its syllable, position of the diphone in its word, left phonetic context and right phonetic context.

The default join cost employs three equally weighted subcomponents for pitch, energy and spectral mismatches. Spectral discontinuity is estimated by calculating the Mahalanobis distance between two vectors of 12 MFCCs from either side of a potential join point, the Mahalanobis distance between two additional coefficients for F_0 and energy are likewise used to estimate the pitch and energy mismatch across the join.

4. Tuning the voices

Tuning the voices consisted of nothing more than inspecting a few F_0 tracks to set cut-off pitch range parameters for the pitch tracker. Previous experience has shown that the majority of quality problems with *multisyn* voices result either from bad labelling or bad pitch tracking. If a speaker has a large number of phone segments which do not contain a pitch mark, then this is a sign that either the pitch marking has not worked very well, or the speaker is speaking very fast and reducing or deleting phones. If the pitch marking is bad, the pitch-synchronous joining algorithm is unable to select appropriate units; if the speaker is speaking fast, the accuracy of the labelling decreases because the predicted label sequence is less likely to match the actual speech. Either way, the quality of the resulting voice is reduced. There were a few potential problems with some of the Blizzard voices and a small number of diphones were being removed from the final inventories because of this.

	conv	guten	mrt	news	sus
unilex lexicon	1	1	0	24	9
cmu lexicon	0	0	2	18	16

Table 1: Numbers of words from sections of the test set that were missing from our pronunciation lexica.

5. Generating the test sentences

The test sentences were automatically synthesised from the supplied festvox description files. Our first impression on listening to the test set for each voice was that these voices did not sound very good. There were clearly bad joins in many of the examples, which is probably due to the dataset not being big enough to provide a suitable distribution of units in different contexts – this affects the intelligibility of the voices. There were also noticeable problems with the intonation of many of the sentences. This is not too surprising because the type of voice built has no intonation model or pitch modification. Instead, the *multisyn* engine assumes that, by choosing units from a suitable context, a natural intonation contour will also be generated. For small datasets, like the Blizzard voices, many contexts will be missing; units selected from the wrong context will result in bad intonation. We noticed a particular problems with the `rms` voices. The original `rms` data has a low speech rate. Here, bad intonation is particularly noticeable, and even sounds exaggerated. However, the evaluation result, suggest that listeners did not have a problem with this.

This finding agreed with our previous experience that the ARCTIC datasets are not ideal for building *multisyn* voices, see section 6 for more on this.

It was also noted that there were a reasonable number of out-of-lexicon words in the test set. Because the test sentences were "unseen", we decided not to add the missing words to the lexicon. Table 1 shows the distribution of words missing from the lexica used. Most of the missing words from the news section are names while most of the missing words from the semantically unpredictable sentences are very low frequency words or compound nouns.

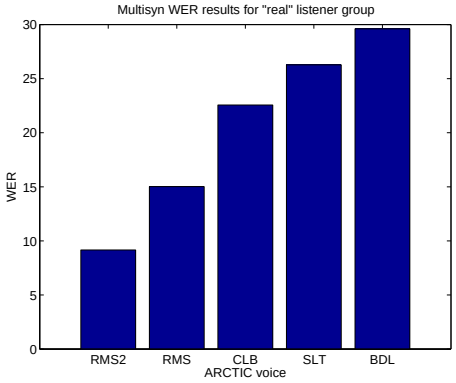

Figure 1: Comparison of word error rates (WER) for the ARTIC voices built for *multisyn*, as judged by the "real" people group. Two voices built from the rms dataset were submitted: RMS, built using the CMU lexicon, and RMS2, built using the Unilex lexicon.

6. Discussion

6.1. Pronunciation lexica

The two voices built from the rms dataset were both submitted for testing. We know that our unilex lexicon is more consistent and complete than the CMU lexicon, and previous experience has shown it generally performs better. However, we presume that the CMU lexicon was used in the ARCTIC text selection process to estimate the required diphone coverage. We generally expect our unilex lexicon to produce better results, although the differences between it and the CMU lexicon would be smallest for the standard US English accents of the Blizzard data, compared to other accents.

The evaluation results from the two voices where quite similar, with one marked difference: the word error rate for subjects that were not speech experts or undergraduate students was substantially lower for the rms voice built with the unilex lexicon (RMS2) than for other voices, including RMS. The results are shown in Figure 1. However, this difference did not show up elsewhere in the results.

To further judge the quality of the ARCTIC based voices, we took a closer look at the difference between the two rms voices and compared the ARCTIC voices to a voice built from a larger dataset. To do this we ran two tests. The first was to compare the two versions of the rms voice using the Blizzard test sentences. The second was to compare the number of missing diphones in RMS2 to those missing in one of our own voices, by synthesising a large test set.

6.2. Comparison of the two rms voices

In this test, we synthesised versions of the supplied test sentences for each voice. Any diphones that were required, but that were not available in a voice, were noted.

We found that the RMS voice (built with the CMU lexicon) only had 4 such missing diphones, suggesting that the ARCTIC dataset contains at least one of each diphone that is likely to occur in sentences using words and pronunciations from the CMU lexicon.

However, when we looked at the RMS2 voice (built with our own lexicon), we found that there were 56 occurrences (46 unique types) of such missing diphones (approximately 1 missing diphone for every 5 utterances). 25 of these were in the news portion of the test set and 14 were in the semantically unpredictable sentences (21 and 13 unique types respectively). This result confirms our hypothesis that using the same lexicon for text selection and at runtime makes a major difference to diphone coverage.

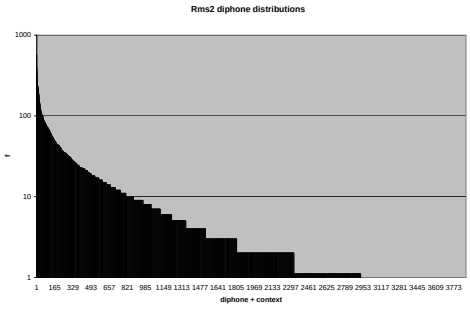

Figure 2: Distribution of context dependent diphones for the RMS2 ARCTIC voice.

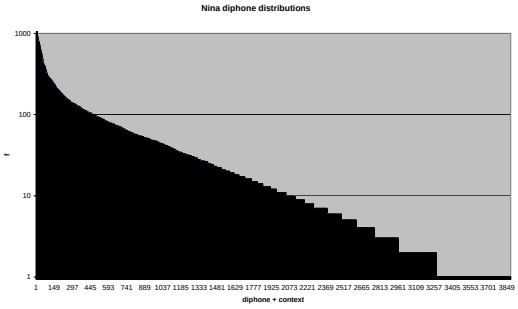

Figure 3: Distribution of context dependent diphones for our own nina voice.

Missing diphones can occur for two main reasons in the *multisyn* build process. Firstly, the diphones could be missing from the recorded speech data due to bad planning or if the speaker deviated from the prompt text or used an unexpected pronunciation. Secondly, the diphone could be present in the speech data yet have been explicitly excluded due to bad labelling or bad pitch marking.

The difference between RMS and RMS2 cannot be explained by differences in pitch marking, since this was identical for the two voices. Our unilex has a larger phone set (55) than the CMU lexicon (47). So it is to be expected that coverage determined by the CMU lexicon may not be sufficient when using the unilex lexicon. To investigate this issue further the unilex version of rms (RMS2) was compared to a dataset (*nina*) designed to be used with the unilex lexicon.

The nina voice used in this comparison is a British English voice built from a larger dataset than the ARCTIC datasets (approximately 175 000 phones compared to approximately 36 000 phones). With this in mind, we first compared the number of distinct diphone types between the voices.We would expect the distributions to be similar as both voices are designed to provide diphone coverage. We then compared the distribution of diphones in different specific contexts, as it is only when there are instances of a diphone available in each different context, that the *multisyn* engine will be able to produce natural prosody without signal processing.

The number of distinct phone types, diphone types and context-dependent diphone types in each voice are shown in table 2 Context-dependent diphones are described by their lexical stress and position with respect to the syllabic structure. This is a simplified version of the context actually used by our text selection process [6].

	phone types	diphone types	CD-diphone types
RMS2	56	1559	2922
nina	51	1851	3880

Table 2: Number of phone types, diphone types and context dependent diphone types in nina and RMS2.

The first point to note is that the nina voice has fewer phone types than the RMS2 voice (because of the different dialect of English). The nina voice contains more diphone types than the RMS2 voice, suggesting that the RMS2 voice may be lacking in basic diphone coverage.

As expected, the nina voice also has more context-dependent diphone types than the RMS2 voice, and a slightly higher ratio of context-dependent diphones to diphones (meaning that each diphone appears in more different contexts). Here, the larger dataset of nina clearly provides a better selection of diphones in different contexts than the `rms` dataset.

Figures 2 and 3 show a plot of frequencies of the context dependent diphone types, sorted with the most frequent first. Plotted on a log scale, the latter part of the nina plot shows how the text selection process successfully raises the number of instances of the less frequent diphones in each of the contexts – in other words, trying to avoid a long tail of zero or low frequency diphones.

To further assess the extent of missing diphones from each of these voices, approximately eight hours of speech from newspaper sentences was synthesised using each voice, and a count made of the number of missing diphones requested. With respect to this dataset, the nina voice was found to have 114 unique missing diphones, whereas the RMS2 voice was found to have 255. These figures are for diphone types; the figures for context dependent diphone types are likely to be even more striking.

7. Conclusions

Our first conclusion is that participating in the Blizzard evaluation has been a useful experience and that we have learnt more about our own synthesis technique through taking part.

Even though we feel that the ARCTIC datasets are not ideal for building *multisyn* voices, the resulting voices are reasonably intelligible, although there is a wide variation across the different voices (e.g. the WER results in figure 1). We believe the size and diphone distributions of the ARCTIC datasets are insufficient to make a good *multisyn* voice. It is possible that signal processing techniques could be used to make up for the lack of diphones in sufficiently different contexts, although we think that a larger (maybe only slightly larger) dataset is a better solution (e.g. our own nina voice). With this in mind, we would like to see future challenges including:

- larger databases
- building voices from datasets of varying sizes – to evaluate how voice quality scales with dataset size
- building voices from a subset of a specified size (e.g. 20k, 50k, 100k, 200k diphones) from a very large dataset – to evaluate text-selection algorithms
- other accents of English
- other languages

8. Acknowledgements

We thank all those that have supported Festival over the years, both developers and users. We greatly appreciate the current funding provided by EPSRC (grant GR/R94688/01) and Scottish Enterprise (under the Edinburgh-Stanford Link) which has partially supported some aspects of this research.

Special thanks go to Tina Bennett and Alan Black (and anyone else at CMU) whose effort and hard work made this evaluation happen.

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] Paul Taylor, Alan Black, and Richard Caley, "The architecture of the Festival speech synthesis system," in *Proc. The Third ESCA Workshop in Speech Synthesis*, 1998, pp. 147–151.

[3] 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.

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

[6] Yoko Saikachi, "Building a unit selection voice for Festival," M.Sc. thesis, Department of Linguistics, University of Edinburgh, 2003.