THE CMU ARCTIC SPEECH DATABASES

John Kominek, Alan W Black
{jkominek,awb}@cs.cmu.edu

Language Technologies Institute
Carnegie Mellon University, USA

ABSTRACT

The CMU Arctic databases designed for the purpose of speech synthesis research. These single speaker speech databases have been carefully recorded under studio conditions and consist of approximately 1200 phonetically balanced English utterances. In addition to wavefiles, the databases provide complete support for the Festival Speech Synthesis System, including pre-built voices that may be used as is. The entire package is distributed as free software, without restriction on commercial or non-commercial use.

1. INTRODUCTION

CMU ARCTIC is a set of studio recorded single speaker databases created with the goal of supporting speech synthesis research. An Arctic database is a reading of the Arctic prompt set (of about 1200 utterances) by a speaker in a specified style of delivery. The audio and EGG recordings are packaged with phonetic labels, pitchmark files, and everything else required to build a Festival unit selection voice [1]. HTS voices are also supported [2].

The first public release of Arctic – version 0.95 – was timed with SSW-5 (this workshop) in mind. This version contains recordings by four separate speakers. Two additional databases are currently under preparation. CMU ARCTIC is released as free software [3]. Licensing conditions are consistent with Carnegie Mellon's Sphinx family of speech software.

2. PRIOR WORK

A widely used, and classic, speech database is TIMIT. This corpus (originating in 1986) was collected to support the training and testing of automatic speech recognition systems. The original distribution is a diverse corpus of American English with 630 separate speakers reading 10 sentences each. In 1997 a freely available, single-speaker version was released by the University of Edinburgh.

Though sometimes described as phonetically balanced, TIMIT is better thought of as phonetically compact. The core 450 sentences of this corpus are not representative of regular English, including sentences that are difficult for non-native speakers to read. Because the phoneme sequences of this database are often unusual, experience has shown that TIMIT is not well suited for synthesis.

TIMIT does have, in addition, a more phonetically diverse prompt set of 1890 sentences, but we are aware of no single-speaker version. After considering recording these, we opted instead to introduce a new speech corpus that better suits the requirements of speech synthesis.

3. DESIGN CONSIDERATIONS

We consider a database good if it:

1. Is readily recordable.
2. Is consistently and cleanly recorded.
3. Suites the underlying synthesis technology.
4. Matches and covers the intended domain.

Our design decisions have been guided by the needs of building English unit selection voices using phoneme sized units. Although there is a trend toward employing very large databases of speech with natural coverage, in the near term it is more tractable to design databases that are relatively small. This makes it easier to release multiple versions by multiple speakers, or multiple versions by the same speaker – thereby enabling a larger variety of voices to be built and studied. Arctic can be recorded in a single day.

In gathering material for the Arctic prompt we chose to use out-of-copyright books available from Project Gutenberg [4]. From here we hand selected 37 short stories and novels written a style that is recognizably modern. As such, Arctic is predisposed towards fictional story reading, one of our target domains of interest.

4. AUTOMATIC PROMPT SELECTION

Beginning with initial material of 2.5 million words of plain text, this was converted it into 168 thousand utterances by using the Festvox script *text2utts*. For the sake of recording we reduced it down to a list of 52 thousand “nice” utterances. By nice we mean sentences or phrases that are easily read by a native English speaking voice talent. In practice this means avoiding unusual and out-of-dictionary words, and restricting prompts to between 5 and 20 words long.

Next the Festvox *dataset_select* script was run to find a compact subset utterances containing at least one occurrence of every diphone. Diphones differing in vowel stress, but are otherwise identical, were considered distinct. This process yielded a prompt list with 668 items. Then *dataset_select* was run another time with the first list removed. This second list contained 629 prompt. The idea behind this repetition this was to give Arctic a built-in redundancy of diphone coverage, but not so much that it became burdensome to record.

Next we attempted trial recordings of the prompts. Various considerations led us to trim the list down to a total of 1132 prompts. This is the version 0.95 list.

Our selection criteria of diphone coverage matches that of concatenative synthesizers performing joins at diphone boundaries, and one should be aware of this inherent bias. Opting for larger base units (such as triphones or syllables) would result in a larger prompt list. The Arctic corpus does not preclude other technologies, however.

5. PHONETIC COVERAGE

The exact definition of phoneme set is a fundamental issue in speech synthesis, and determines what is meant by "complete coverage." There is no easy answer. Festival's lexicon for American English starts with the 40 basic units found in CMU-DICT [6], splits the vowels according to two levels of stress, and adds the reduced vowel schwa and silence symbol 'pau' for a total of 56.

Dict	Lexical Entries	Num Phones	Diphone Counts			
			Intern	Cross	Limit	%
cmu 0.6	127,073	40	1383	1574	1599	98.4
		70	2969	4562	4899	93.1
fest 0.6	112,113	41	1385	1655	1680	98.5
		56	2179	2889	3135	92.2

Table 1. Statistics of the dictionary used in Festival and the current version of CMU-DICT. For each: top row – unstressed phone; bottom row – stressed phones. Column counts are: number of internal diphones, adding in cross-word diphones, combinatorial upper limit, Cross/Limit percent. CMU-DICT has three levels of stress while Fest-Dict distinguishes only two.

Corpus	Prompts	Total Phones	Unstressed Diphone	Stressed Diphone
TIMIT-sx	450	15321	1210	1427
TIMIT-si	1890	72429	1212	1547
TIMIT-all	2342	87819	1312	1692
Arctic-All	168,443	9,541,895	1515	2302
Arctic-A	593	20677	1313	1782
Arctic-B	539	18476	1267	1667
Arctic-C	73	2964	735	825
Arctic	1132	39153	1339	1841
Arctic ABC	1205	42118	1363	1916

Table 2. Diphone coverage of Arctic and TIMIT, using Festival phoneset and utterance transcriptions.

Table 1 summarizes the number of diphones that are possible using Festival's lexicon: 1655 when ignoring stress, 2889 including stress. The prohibited (unstressed) diphones all have NG as the second phone. Table 2 compares diphone coverage of various Arctic and TIMIT prompt lists. The full corpus of 168K utterances contains 1532 unique diphones – 123 short of the upper limit.

Though the coverage of Arctic is larger than TIMIT (1339 vs. 1312), due to various stages of filtering it falls short of being complete. This is an admitted defect. The 73 utterances of Arctic-C is a tentative (i.e. pre-release) prompt list intended to fill the worst of the gap. It add 24 unstressed and 75 stressed diphones to the mix.

Filling more holes becomes increasingly laborious, as the elusive diphones tend to embed themselves in utterances that are not "nice" to record. More fundamentally, at some point a noise floor is reached. The combined effects of lexicon irregularity, speaker variation, and labeling errors casts question onto the very validity of these rare diphones. Beyond a certain threshold, it is doubtful whether rare diphones should be collected and labeled at all, instead relying on a backoff mechanism when they occur in text.

6. CONCLUSION

The current release of CMU ARCTIC contains 1 female and 3 male voices. This variety is useful when reading material – such as children's stories – that shifts between narration and talking character [6]. To support this and other research interests, additional voice databases will be compiled and released over time.

7. ACKNOWLEDGMENTS

This material is partly based upon work supported by the U.S. National Science Foundation under Grant No. 0219687, "Evaluation and Personalization of Synthetic Voices." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

8. REFERENCES

[1] A. Black, K. Lenzo, "Building voices in the Festival speech synthesis system," 2000, http://festvox.org/bsv.

[2] K. Tokuda, H. Zen, A. Black, An HMM-based Approach to English Speech Synthesis," Proc of Acoustical Society of Japan, 3-10-15, Sept. 2002.

[3] J. Kominek and A. Black, "The CMU ARCTIC databases for speech synthesis," Tech. Rep. CMU-LTI-03-177, Language Technologies Institute, Carnegie Mellon University, 2003. http://www.festvox.org/cmu_arctic.

[4] M. Hart, Project Gutenberg, 2003, http://promo.net/pg.

[5] Carnegie Mellon University, "The CMU pronunciation dictionary", 2000, http://www.speech.cs.cmu.edu.

[6] J. Zhang, A. Black, R. Sproat, "Identifying Speakers in Children's Stories for Speech Synthesis," Eurospeech 2003.