CMU ARCTIC
databases for speech synthesis

John Kominek and Alan W Black

CMU-LTI-03-177

Ver. 0.95

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Abstract

This report introduces 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 nearly 1150 phonetically balanced English utterances. They are distributed as free software, without restriction on commercial or non-commercial use.

The Arctic corpus consists of four primary sets of recordings (3 male, 1 female), plus several ancillary databases. Each database is distributed with automatically segmented phonetic labels. These extra files were derived using the standard voice building scripts of the Festvox system. In addition to phonetic labels, the databases provide complete support for the Festival Speech Synthesis System, including pre-built voices that may be used as is.

Festival and Festvox are available at http://www.festvox.org.

The Arctic speech corpus is available at http://www.festvox.org/cmu_arctic.

1 Introduction

The idea of a having common set of resources targeted towards the needs of speech synthesis research has been discussed for many years, but only partially fulfilled. Systems such as Festival [2] and MBROLA [7] – by organizing and documenting many of the necessary algorithms – have made it significantly easier for people to enter the field of speech synthesis. However, the supply of publicly available speech databases is small, and has lagged behind the needs of current technology. The development and release of CMU Arctic is intended to address this shortcoming.

CMU Arctic is a set of single speaker databases that have been carefully recorded under studio conditions, packaged with associated information such as phonetic labels and pitchmark files. An Arctic "database" is a reading of the Arctic prompt set (plus associated files) by a single speaker in a specified style of delivery. This release of Arctic contains recordings by four separate speakers. When referring to the Arctic "corpus" we mean the entire collection of databases, including test sets.

The databases have version numbers. As with computer code, version numbers indicate the level of maturity and stability. Numbers with a zero after the decimal point (e.g. version 1.0) are major releases intended to serve as a reference point for system development and evaluation. Minor releases are subject to change, allowing for more frequent additions, deletions, and improvements. All releases fall under a BSD-style open software license, described in Appendix A. This document reports on version 0.95.

2 Speech Databases

The majority of existing databases have been prepared primarily with automatic speech recognition in mind. Prominent examples include TIDIGITS [15] (isolated word recognition), SWITCHBOARD [10] and CALLHOME [4] (spontaneous phone conversations), and Aurora [12] (noisy speech). Databases that are designed for training and testing of ASR systems require large amounts of speech collected under realistic and noisy conditions, by multiple speakers with broadly varying accents. These characteristics are not well suited for constructing synthetic voices. Building high quality synthetic voices requires a much greater degree of control, since the flavor of the voice invariably reflects the nature of the recordings.

For a speech database to serve as the basis for constructing a synthetic voice, the recordings should be of studio quality and free of noise. Noise includes not just external sounds such as fans and squeaking chairs, but unwanted breaths and clicks. Also, *what* is recorded matters a great deal. Since perfect quality open-domain synthesis is not yet possible, the recorded utterances need to reflect the target domain – in particular, by being phonetically balanced. Finally, the prosody of speech needs to be controlled so that the synthetic voice's style of delivery is both consistent and appropriate. Satisfying these requirements makes a corpus *designed* for synthesis, as opposed to merely collected.

Probably the most common resource for speech synthesis research is Boston University's FM Radio News Corpus [16]. This corpus (recorded in 1994) is now almost ten years old. It consists of seven professional radio announcers reading either pre-edited or off-the-wire news stories. As such, the recordings are well suited for a study of prosody in speech – the primary intention of this corpus. However, the Boston Corpus predates the advent of unit selection synthesis, the dominant technique of the past decade and our principal research interest. For this purpose the Boston recordings contain too little recorded speech for the amount of prosodic effect contained. Also, the voices often have an unwanted creaky quality and an excess of breathiness.

An even older database is TIMIT [9]. This corpus (recorded in 1986) was collected to support the training and testing of automatic speech recognition systems. The original NIST distribution is a diverse corpus of American English with 630 separate speakers reading 10 sentences each. Though sometimes described as phonetically balanced, it is better thought of as phonetically compact. The core 450 sentences of this corpus are not representative of regular English and include many (near) tongue twisters that are difficult for non-native speakers to read. TIMIT does have, in addition, a more phonetically diverse prompt set of 1890 sentences, but we are aware of no single-speaker version of these.

In 1997 a freely available, single-speaker version of the TIMIT prompt set was released for synthesis research by the University of Edinburgh [8]. But because the phoneme sequences of this database are unusual, experience has shown that TIMIT-based voices tend to be sub-par. Such experience has encouraged us to construct larger and better databases.

Each Arctic database consists of nearly 1150 utterances, most being between one and four seconds long. The prompt list is split into two sets (A and B), each of which is designed to be phonetically balanced American English and have diphone coverage representative of the source material. The wavefiles were recorded in a sound proof booth at 32,000 Hz with simultaneous EGG (laryngograph) measurements. This release includes recordings from four voice talents: male General American US English, female General American US English, male Canadian accented English, and male Scottish accented English. In all cases the lexical and phonetic descriptions derive from the US English front-end module distributed with Festival. In this configuration Festival employs CMUDICT [5] as its dictionary component. Thus the two accented databases are described using a General American phoneme set and lexicon, despite any speaker-specific deviation.

2.1 Source Text

Since it is very important to us that use of the Arctic corpus be unrestricted, we needed to start from a source of written material that does not impose any copyright restrictions incompatible with our aims. Although there are legally defined "fair use" rights in the US that specifically allow for the extraction of short quotes from a larger body of work, such rulings principally consider the needs of scholarship and of review (for which specific attribution is apparent). Our use does not exactly fit this category, causing us to be cautious. We don't want there to be any residual questions about the availability of this release.

Because we are releasing Arctic under a "free software" license that explicitly allows for commercial exploitation in addition to university research, it is not enough to reside behind a license that permits "for research use only." Thus we decided to select our sentences from the largest text corpus available that has compatible copyright restrictions – the Gutenberg Project [11]. The Gutenberg Project aims to collect and publish online all out-of-copyright books of the English language. Technically, the Gutenberg license is a free software license applied to written text. It includes a clause that allows their particular copyright (which is affixed to the work itself) to be removed and have the remaining text enter the public domain.

To begin our data collection we selected a portion of Gutenberg books and extracted the text body proper, discarding the surrounding legal matter. We did this not to redistribute the texts absent of or under different copyright, but simply to avoid recording sentences written in "legalese."

The details of prompt design and extraction are described in the section that follows.

3 Prompt Design

From our experience in unit selection synthesis we are very much aware that good speech databases make for better voices. We consider a database good if it:

1. Is readily recordable.
2. Suites the underlying synthesis technology.
3. Matches the intended domain.

Our design decisions have been guided by the needs of building English unit selection voices operating with phoneme sized units. Although there is a trend toward employing very large databases of speech with natural coverage [13], 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, thereby enabling a larger variety of voices to be built and studied. Arctic can be recorded and quality-verified in a single day. Typically, the voice talent will record Set A in the morning and Set B in the afternoon.

As anyone who has built a unit selection synthesizer knows, the quality of output is highly dependent on the coverage of the database. Achieving coverage is fairly straightforward for limited domains but extremely difficult for others. For example, (short of recording all stage performances in total), achieving full coverage of Shakespeare's oeuvre of plays – with convincing emotional delivery – would be extremely difficult in the least. For the Arctic project we have chosen as our target domain fictional prose; in particular, short stories that have a modest amount of dialog and can be narrated from a single perspective. We believe that this choice is both challenging and important, and yet within reach of current technology.

Designing the Arctic prompt set followed seven stages.

1. Decide on a target technology. (Unit selection synthesis)
2. Decide on the target domain. (Short stories)
3. Select a document source. (Project Gutenberg)
4. Select source documents.
5. Automatically select select sentences from the source text.
6. Inspect and remove unsuitable sentences.
7. Perform a trial recording and prune out difficult utterances.

As described in section 2.1, we chose to use out-of-copyright books from the Gutenberg Project. With most of these texts being at least 70 years old, we face the issue of language drift. The English language has changed considerably over the past centuries and we did not want to infuse in our prompt set archaic English sentences. Thus we have hand selected a set of short stories whose style is recognizably modern, if not completely contemporary. Partly for consistency and partly from personal preference, we selected stories largely from the early 20th century author

Jack London. Many of these stories – famously "To Build a Fire" – depict the difficult living conditions of the Yukon. Other selected books also describe the far Canadian north, hence our moniker *Arctic*. Appendix B lists the source text files.

3.1 Automatic Prompt Selection

Starting with our initial text corpus of 2.5 million words and 168 thousand utterances, we ran the Festvox [1] script *text2utts*. This gave us a list of 52 thousand "nice" utterances. By nice we mean utterances (sentences or phrases) that are easily read by a native English speaking voice talent. This has two aspects: length and pronounceability.

With respect to length, we filtered out sentences that are not between 5 and 15 words. Short utterances often have a different prosodic delivery than sentences of normal length. Excluding these, though, does mean that synthesizers built from this database are likely to be less than optimal for reading very short phrases. Conversely, sentences longer than 15 words are difficult to read aloud without making a mistake. It is hard enough already to read over a thousand utterances consistently and correctly. Not being especially interested in modeling speech disfluencies, we cannot afford to make the task more strenuous by the inclusion of lengthy sentences.

The second key restriction is that all words of a selected utterance must already be in the lexicon CMUDICT. Although Festival has reasonable letter to sound rules, we wish to reduce the chance of predicting pronunciations differently from how our voice talents actually say the prompts. Restricting sentences to contain only known dictionary words helps reduce (but not completely eliminate) errors of this kind. We did consider including words with only a single pronunciation (i.e. by excluding homographs) but that turns out to be excessively restrictive.

Next the Festvox *dataset_select* script was run to search for the subset of the 52K nice utterances having the best diphone coverage. In order to encourage more thorough coverage we tagged vowels with the stress value (0 or 1) of the syllable in which they are contained. Note that *dataset_select* employs a greedy algorithm and so is unlikely to find the global optimum, but will come close. Reference [3] describes an elaborate method for selection based on coverage of a given large corpora of text and using an explicit modeling of the acoustic-phonetics of a particular speaker. In the construction of CMU Arctic, however, we used the simpler approach encoded in the *dataset_select* script, as it appears sufficient.

At this point 668 utterances had been extracted from the candidate set. These were removed, followed by a second run of *dataset_select*. This resulted in a second set of 629 utterances with good (but not necessarily complete) diphone coverage. This second list is smaller than the first because diphones that appear only once in the corpus have already been extracted during the first pass.

3.2 Further Hand Pruning

The results of automatic selection are still not ideal. We further winnowed the prompt lists in two stages of hand pruning. The first examination is simply based on visual inspection. Criteria for exclusion include: archaic terminology, awkward grammar, confusable homographs, hard to pronounce foreign names, and various embarrassments such as swear words.

Next, the two of us performed trial recordings of the prompt set. From this experience we further removed utterances deemed too hard to record or too liable of mispronunciation. Deciding on the exact cutoff is tricky. Doing too little pruning increases the burden of recording and repair work. Going too far reduces the phonetic coverage below the desired level. In the end, the reduced sets A and B contain 593 and 539 prompts respectively, for a total of 1132.

Finally we normalized punctuation and updated spelling. For instance, reading that "to-morrow comes after to-day" undeniably stalls the modern eye. It helps to lowercase words that happened to sneak in as ALL CAPS, and to reduce question marks and exclamation points down to periods. Utterances should resemble declarative sentences in that they begin with a capital letter and end with a period. All these alterations help to deliver the prompts under consistent control.

The Arctic prompt set is not perfect but does achieve our objectives. There do remain utterances that are unusual and can trip up a voice talent[1] but none are truly awkward[2]. Partial pronunciation guidance is found in Appendix C.

Table 1. Number of utterances through three stages of filtering

Arctic	*Automatic*	*Hand Pruning*	
	Stage	*Pass1*	*Pass2*
Set A	668	597	593
Set B	629	541	539
Total	1297	1138	1132

3.3 Phonetic Coverage

As explained in Section 3.1, one of the criteria used in the design of the Arctic prompt set is that it exhibit good diphone coverage with a minimal amount of text. A comparison with other corpora reveals the advantage of this approach.

In Table 2 we see that the 1132 prompts of Arctic translate into 39,153 phonemes. This figure comes from the front-end element of Festival that converts prompts into phonemic label files, and includes silence phones at the beginning and

[1] arctic_a0508: "Soon shall it be thrust back from off prostrate humanity."

[2] From TIMIT: "Will a robin wear a yellow lily?" and "A roll of wire lay near the wall."

ending of each utterance. For this tabulation the phoneme set has 41 elements consisting of the 39 phonemes from CMUDICT, plus the reduced vowel schwa /AX/ and the pause symbol /PAU/. The percentages for diphone and triphone coverage are based on simple combinatorics, not on an exhaustive list of n-gram phoneme sequences that are realizable in English. Thus the number of possible diphones is 1680=41x41-1.[3] Arctic achieves nearly 80% and 14% respectively. This diphone coverage is significantly higher than that of the Boston University Radio news F2b corpus [16], and is higher even than that of the entire 2342 prompt TIMIT corpus [9]. TIMIT does offer greater triphone coverage though, due to the larger amount of text contained.

Some of the entries in Table 2 warrant a brief mention. Kal-Text4 is an older (and smaller) database that we have used in previous work [14]. It was designed along similar principles as Arctic and prior experience with it has informed our current effort. The corpus Uniphone is our minimal prompt list for achieving full phoneme coverage of English. In contrast to TIMIT-SA, each unit appears once and only once.

Uniphone

1. “A whole joy was reaping.”
2. “But they've gone south.”
3. “You should fetch azure mike.”

TIMIT-SA

1. “She had your dark suit in greasy wash water all year.”
2. 'Don't ask me to carry an oily rag like that.”

Table 2. Coverage of various corpora. The number of unique words found in a corpus are listed in the “Unique” column. Refer to main text for description of Arctic “All Utts” and “Nice Utts”. A small number (1.8%) of nice-utts diphones did not make it into Arctic.

Corpus	**Total Number of Units**				**Coverage**		
	Prompts	Words	Unique	Phones	Phoneme	Diphone	Triphone
Uniphone	3	14	14	46	100%	2.56%	0.06%
TIMIT-sa	2	21	21	69	100%	3.87%	0.09%
TIMIT-sx	450	3397	1184	15321	100%	72.2%	9.1%
TIMIT-si	1890	17343	5516	72429	100%	72.1%	17.4%
TIMIT-all	2342	20771	6614	87819	100%	78.2%	19.4%
BUR-f2b	155	8726	2758	39470	100%	65.2%	11.6%
Kal-Text4	534	4000	1553	14905	100%	61.1%	8.4%
All Utts	168,443	2,545,221	49,126	9,541,969	100%	90.2%	43.6%
Nice Utts	52,186	495,790	19,948	1,827,355	100%	81.4%	33.0%
Arctic-A	593	5284	1958	20677	100%	78.0%	10.2%
Arctic-B	539	4761	1775	18476	100%	75.4%	9.7%
Arctic	1132	10045	2974	39153	100%	79.6%	13.7%

[3] This tally is reduced by one because /pau-pau/ does not count as a valid diphone.

The graph below shows a histogram of Arctic phoneme frequencies. The most common phonemes (besides silence/pause) are /N/ and /AX/. The least common are /OY/ and /ZH/.

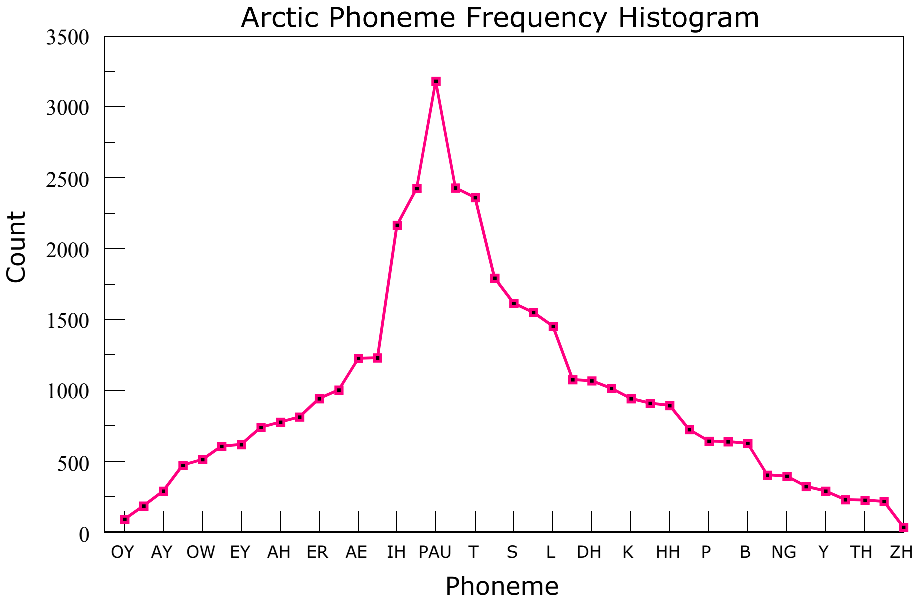

Figure 1. Arctic distribution of phonemes. The /PAU/ phone includes both bracketing silences and internal pauses. The entropy of this curve is 4.998, compared to log_2 (42) =5.392 for a completely flat distribution of 42 symbols.

4 Recording Conditions

The Arctic databases have been recorded in a soundproof booth located in CMU's speech lab. The booth[4] is a double steel walled chamber with a six inch air gap between the inner and outer chambers, and has foam baffling mounted on the inside walls to damp out resonances. The voice talent being recorded sits with their mouth 6 to 12 inches from a Sennheiser MD431 near field condenser microphone. Optionally, a pop filter may be installed between the speaker and the microphone to reduce the force of air puffs emerging from bilabial plosives ([p^h] and [b^h]) and other strongly released stops, although this was not done for release 0.95. A high fidelity capture card[5] performed dual channel analog-to-digital conversion at 32,000 Hz.

In addition to the microphone setup, the speaker secures to their larynx a pair of electrodes from an Electro-Glottal Graph machine[6]. This allows us to directly record activity of the glottis during voiced segments. Figure 2 below shows an example waveform and its relation to the corresponding speech signal.

[4] Manufactured by Acoustical Solutions Corp.

[5] Model CardDeluxe, manufactured by Digital Audio.

[6] Model EG2-PC, manufactured by Glottal Enterprises.

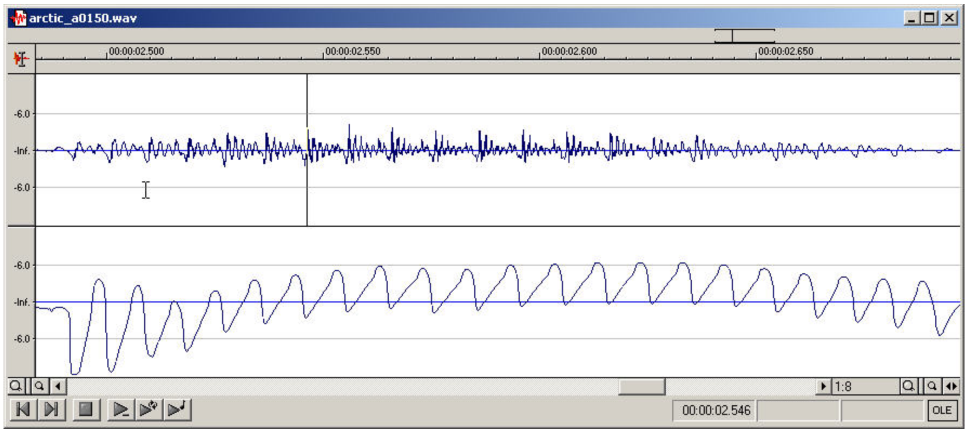

Figure 2. Stereo signal showing speech and EGG waveforms.

Two possible uses of EGG signals are to more accurately detect the onset and disappearance of voicing, and to improve the fidelity of pitchmark detection. Care must be taken, however, to account for the delay between channels.

Because in any large prompting of speech some utterances are going to be miscues, not all of the recordings are from the same session. Talents are instructed to speak in a flat voice with minimal inflection. This helps reduce the unevenness that results from recording over multiple days. When discovered, broken recordings are removed from the database and queued up for re-recording. Such repair work also necessitates a rebuilding of the base Festival voice and associated files, and so can lag behind an initial release. This is one reason why Arctic databases have version numbers.

5 Labeling and Voice Building

As these databases are designed for speech synthesis research, we wish to provide basic annotation that we know is sufficient to allow research with this work. Second, we wish to provide a baseline reference against which others (and even ourselves) can demonstrate improvements.

Thus we used the Festvox voice building tools [1] to build a complete unit selection voice for each of the recorded sets. Importantly, this process provides phonetic timing labels and the construction of Festival-style utterance structures, from which durational and other linguistic models may be derived.

The phonetic labeling stage uses CMU SphinxTrain [6] to build full HMM-based acoustic models from the particular speaker's recordings. This works because the databases are phonetically balanced and because the amount of recording is sufficient for speaker-specific models. SphinxTrain then uses these models to perform forced alignment, which yields phoneme beginning and ending times.

In this first release we have done no hand correction of labels, which we believe is proper for a baseline reference. At some later date, when time and resources permit, hand correction of these databases would be a useful addition. Corrected labels not only yield better sounding voices, but serve as a benchmark against which automatic labeling techniques can be compared.

The voice building process exactly follows the standard procedure as described in the Festvox manual. In brief, the steps are:

- build prompts
- run sphinxtrain to label phonemes
- extract pitch marks
- create mcep feature files
- build a clunits voice.

6 Conclusion

The construction and release of the CMU Arctic corpus has been motivated in part by the success experienced during the 1990s in the speech recognition community. The workshops organized by NIST and DARPA saw a sharing of successful techniques and a steady improvement in word error rates. To a large extent this is attributable to the common training and test sets that were collected and made available to researchers.

The same strategy would also serve the speech synthesis community well. It is not without pitfalls, however. The excessive focus on word error rates, it can be argued, led the speech recognition world to dwell on small incremental improvements to established techniques (HMMs), curtailing the opportunity for radically different – and potentially better – approaches to succeed. In speech synthesis, however, we do not have a quality measure that is as easy and reliable to calculate as word error rate, and so there is less danger in developing tunnel vision.

Though the Arctic corpus offers a significant resource for speech synthesis research, it is not all-purpose. Voices built from our data will reflect the source material and original recordings, for better or worse. We expect that they will be good for reading short stories, bad for reading poems, and adequate for dialog systems. Also, the corpus needs the addition of a test set for system evaluation, and would benefit greatly from carefully hand-corrected labels.

The design of CMU Arctic has been guided by the needs of unit selection synthesis but is not constrained to that technology. We feel (and hope) that it will prove useful for HMM-based speech synthesis, articulatory synthesis, prosody analysis, and voice conversion, among other things yet to be devised.

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 National Science Foundation.

We are deeply indebted to Brian Langner and Stephanie Tomko their significant contribution to this project.

A Licensing

The Arctic databases are distributed as “free software” under the following terms.

Carnegie Mellon University
Copyright (c) 2003
All Rights Reserved.

Permission to use, copy, modify, and license this software and its documentation for any purpose, is hereby granted without fee, subject to the following conditions:

1. The code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Any modifications must be clearly marked as such.
3. Original authors' names are not deleted.

THE AUTHORS OF THIS WORK DISCLAIM ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

All voice talents have signed a waiver agreeing to distribution of their recordings under these terms.

B Source Texts

Raw material for the Arctic corpus comprises 43 e-books (text files), 34 of which are from the author Jack London. The five Robert Service books are collections of poems and do not figure in the construction of the 1132 item Arctic prompt list. They are “reserved for future developments.”

The numbers listed in the table are taken from output of the popular Unix command 'wc'.

Author	Filename	Lines	Words	Bytes	Comment
London	badam10.txt	4691	39055	216936	
	bdlit10.txt	12709	112666	644092	
	callw10.txt	3293	31865	182821	
	advnt10.txt	8523	70837	408703	
	cwolf10.txt	12249	106392	591701	
	elsnr10.txt	12861	113564	641166	
	fthmn10.txt	5191	46663	265415	
	gdlgh10.txt	5591	51443	290861	
	hmndr10.txt	3996	31705	184073	
	hsprd10.txt	3346	30348	175213	
	irnhl10.txt	10225	88062	514224	
	jaket10.txt	11276	103709	578160	
	jbarl10.txt	6742	64719	364361	
	jrisl10.txt	7566	69911	402231	
	klndk10.txt	5868	52123	293712	
	llife10.txt	5422	48882	271388	
	lstfc10.txt	4270	41053	228638	
	mcjer10.txt	11184	96494	555859	
	meden10.txt	15422	140148	792268	
	mface11.txt	5653	47953	272420	
	mklmt10.txt	5979	55208	313098	
	ntbrn10.txt	6472	52102	294072	
	smkbl10.txt	5471	41609	237879	
	snwlf11.txt	5346	48553	271446	
	soset10.txt	5118	51037	285406	
	sstrg10.txt	4886	47234	269392	
	tgame10.txt	1670	15286	87413	

Author	**Filename**	**Lines**	**Words**	**Bytes**	**Comment**
	totfp10.txt	3509	31381	174676	
	tpota10.txt	6976	62243	354725	
	tred110.txt	4200	37831	215479	
	vlymn10.txt	20443	167021	941140	
	wrcls10.txt	3573	34229	205500	
	wtfng10.txt	7832	72092	408227	
		245345	2182857	12382038	author subtotal
Curwood	flwnt10.txt	8119	71436	397780	
	nmdnt10.txt	6566	64829	353124	
		14685	136265	750904	author subtotal
Conner	cplcn10.txt	16219	124461	720696	
Hakluyt	nwpas10.txt	5046	50012	283196	
		21265	174473	1003892	author subtotal
Subtotal		281,295	2,493,595	14,136,834	
Service	bchee10.txt	2347	18962	104120	Collections of poems...
	blbhm10.txt	6413	40551	220956	
	redcr10.txt	3746	25707	138598	
	rolst10.txt	3571	23345	127628	
	spyuk11.txt	1902	13535	74231	
		17979	122100	665533	author subtotal
Total		299,274	2,615,695	14,802,367	

C Problem Words

From our experience in recording CMU Arctic we have found a small number of words prone to variation. Almost all are proper nouns. Below are some examples with recommended pronunciation. A problem word may be found in multiple prompts; only the first occurrence is listed.

Occurrence	Word	Gloss	Pronunciation
arctic_a0002	Whittemore	wit-more	W IH T M AO R
arctic_a0066	Jeanne	jean	J IY N
arctic_a0069	Eileen	eye-lean	AY L IY N
arctic_a0260	Junta	hoon-tah	HH UH N T AH
arctic_a0319	Edinburgh	ed-in-burr-oh	EH D AH N B ER OW
arctic_a0431	Wada	wah-dah	W AA D AH

Notes:

1. Voice talent jmk pronounces "Eileen" as /AY L IY N/ (rhymes with eighteen).

D Recording Notes

In any sufficiently long recording of speech some wavefiles are bound to be bad. The table below summarized the dates of original recording and of any repair sessions. For each speaker in the Arctic distribution the actual list of prompts recorded can be found in the file ~/etc/txt.done.data. The voice slt contains all the prompts of the final list.

Voice	***bdl***	***slt***	***jmk***	***awb***
Version	***0.95***	***0.95***	**0.95**	**0.90**
Set A	Jul 24, 2003	Aug 12, 2003	Jul 17, 2003	Jul 10, 2003
Set B	Jul 22, 2003	Aug 12, 2003	Jul 17, 2003	Jul 10, 2003
Repair 1	Jul 23, 2003	None	Aug 1, 2003	None
Repair 2	Aug 1, 2003		Sept 10, 2003	
Prompts	1131	1132	1114	1138
Duration (s)	3058.0	3062.7	3255.9	4777.0

Notes:

1. The awb prompts were recorded at 16K and are mono, without a matching EGG signal. The version number for this database is 0.90.
2. The duration row lists total wavefile length. The bdl, slt, and jmk databases have been tightly trimmed but the awb files contain generous amounts of silence before and after the speech segment.
3. The 18 files of repair session 2 for jmk have not been incorporated into the 0.95 release.

E Voice Talents

All of the voice talents in this release of CMU Arctic speak English as their native language. Two of the four (bdl, slt) speak a north midland General American dialect (often simply called "Midwest"), and so most closely match the phonetic slant of CMUDICT. The accent of jmk represents a minor deviation, while that of awb is strikingly different. None of the speakers are smokers.

Voice	***bdl***	***slt***	***jmk***	***awb***
Gender	male	female	male	male
Age	23	31	38	41
Height	6'0" (183)	5'6" (168)	5'10" (178)	5'11" (180)
Education	BS	MS	MS	PhD
1st Language	English	English	English	English
Dialect	North midland American	North midland American	Ontario Canadian (Southern)	South Eastern Scottish (Edinburgh)

References

1. A. W. Black, K. Lenzo, Building voices in the Festival speech synthesis system, 2000, http://festvox.org/bsv.
2. A. W. Black, P. Taylor, R. Caley, The Festival speech synthesis system, 1998, http://festvox.org/festival.
3. A. W. Black, K. Lenzo, Optimal data selection for unit selection synthesis, 1998.
4. A. Canavan, D. Fraff, G. Zipperlen, CALLHOME: American English Speech, 1997.
5. Carnegie Mellon University, The CMU pronunciation dictionary, 2000, http://www.speech.cs.cmu.edu.
6. Carnegie Mellon University, SphinxTrain: building acoustic models for CMU Sphinx, 2001, http://www.speech.cs.cmu.edu.
7. T. Dutoit, V. Pagel, N. Pierret, O. van der Vreken, F. Bataille , The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes, 1996, http://tcts.fpms.ac.be/synthesis/mbrola.html.
8. University of Edinburgh, Center for Speech Technology Research, CSTR US KED TIMIT, 2002, http://festvox.org/dbs/dbs_kdt.html.
9. W. Fisher, D. Doddington, K. Goudie-Marshall, The DARPA speech recognition research database: specifications and status, 1986.
10. J. Godfrey, SWITCHBOARD: Telephone speech corpus for research and development,, 1992.
11. M. Hart, Project Gutenberg, 2003, http://promo.net/pg.
12. H. Hirsch, D. Pearce, The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions, 2000.
13. H. Kawai, M. Tsuzak, A study of time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis, 2002.
14. J. Kominek, T. Bennett, A. W. Black, Evaluating and correcting phoneme segmentation for unit selection synthesis, 2003.
15. R. Leonard, A database for speaker-independent digit recognition, 1986.
16. M. Ostendorf, P. Price, S. Shattuck-Hufnagel, Technical Report ECS-95-001. The Boston University Radio News Corpus, 1996.