1 Articulation Index Corpus©

This file is pcjba:/home/jba/LDC/DOC/doc/LDC.tex—pdf Please see the LDC readme.txt for a rough sketch of the DVD contents.

Identification of the database: Author: Jonathan Wright
Name: Articulation Index Speech V1.0
Number: LDC2004E1

1.1 readme.txt

Articulation Index Corpus[1]

The contents of this DVD are as follows:

readme.txt this file

doc/ directory for documentation (text files)
- doc/doc.txt primary corpus documentation
- doc/darpabet.txt describes the "darpabet" phonetic representation
- doc/speaker_inf.txt speaker demographic information
- doc/speaker_doc.txt describes format of "spkr.txt"
- doc/carrier.txt carrier phrase word sets
- doc/inv_doc.txt description of tables in "inv" directory

inv/ directory of inventory files (text files)

prompts/ directory of prompt files (text files)

syls/ directory for syllable audio data (sphere files)
- syls/wb/p wide-band (16 KHz, 16 bit pcm) phrase data
- syls/wb/s wide-band syllable data
- syls/nb/p narrow-band (8 KHz, 8 bit ulaw) phrase data
- syls/nb/s narrow-band syllable data

All narrow-band data has been "shorten" compressed.

The four sub-paths under syls each contain a set of 20 subdirectories, representing the 20 speakers in the corpus. Each speaker directory contains approximately 2000 sphere files: *(4 data types * 20 speakers * 2000 files/speaker)* representing a total of about 160,000 speech files under syls/.

Please see doc/doc.txt for further information, including how to interpret the file names.

[1]The *Articulation Index Corpus* is copyrighted by LDC, Univ. of Penn., Phila., PA

e	ai	bait	@	a	hat	i	ee	bee	E	e	bet
I	i	hit	o	oa	boat	O	oy	boy	a	o	cot
u	oo	boo	U	u	put	A	u	hut	Y	y	why
W	ow	how				n	n	man	m	m	man
b	b	bee	C	ch	church	d	d	dog	f	f	fish
g	g	dog	G	ng	sing	h	h	he	J	j	judge
k	c	cat	l	l	look	p	p	pen	c	au	caught
r	r	real	R	ir	bird	S	sh	she	s	s	see
Z	s	pleasure	T	th	think	D	th	this	t	t	cat
v	v	vow	w	w	win	y	y	you	z	z	zoo

Table 1: *This table lists the "darpabet" symbols defined in "doc/darpabet.txt," used to represent the sounds of English used in the LDC corpus. The first column contains the character which represents a sound, the second column shows which letters in the word correspond to the sound in question, while the third column contains an example English word containing that sound. The sounds are organized into 13 vowels (a, e, i, o, u, y, w), 2 nasal consonants (m, n) and 24 other consonants.*

1.2 doc/doc.txt

The Articulation Index Corpus was partly inspired by the work of Harvey Fletcher, who did a number of perceptual experiments involving English syllables during the first half of the 20th century. His term "articulation index" meant something like "perceptual index of syllables" where those syllables weren't necessarily words, and reflected how well speakers could correctly identify syllables in the presence of noise. This corpus was created to facilitate similar experiments, as well as to potentially facilitate new methods in speech recognition research.

The basic concept behind the corpus is to record speakers pronouncing syllables of English, some of which might be real words, but most of which are nonsense syllables. The goal was to have each speaker say a set of 2000 syllables common to all speakers, as well as a set of 20 syllables unique to that speaker. This goal was nearly met, but not precisely; see below for a description of the syllable inventory.

1.3 Syllable Selection

The darpabet was chosen as the representation for syllables; doc/darpabet.txt describes the subset of darpabet used in this corpus.

The syllables were selected in two ways. First, all diphone (CV, VC) syllables which were considered valid English syllables were included in the common set (see below about "validity"). These syllables accounted for over 600 of the 2000. Second, the remaining syllables in the common set, as well as those in the speaker unique sets, were chosen based on a word frequency table of Switchboard 1. Phonetic representations for the words were found by lookup in PRONLEX, and then the word frequency was added to the overall syllable frequency for each component syllable of the word. The syllables for the common set and the unique sets were then chosen from the top of this list (the common set coming from higher in the list than the unique sets). This means that, while English syllable frequency might

be difficult to define or calculate, the selection of syllables for this corpus does correlate to some significant extent with the observed frequency of syllables in casual conversations (as represented in Switchboard 1). These syllables chosen based on frequency are all triphones (CVC, CCV, VCC).

Obviously "valid syllable" is a debatable concept. I won't go into full detail about what I considered valid, but two things in particular should be noted about my choice of syllables: syllables with schwa as their nucleus were not selected, but syllables with syllabic "r" as their nucleus were selected. The exclusion of schwa was based on it's degenerate phonological status; that is, it's not a distinctive vowel in English, and it's distribution is based on stress, which wasn't varied in this corpus. Syllabic "r" on the other hand is arguably a true vowel, based on its distribution and phonetics.

1.4 Recording Procedure

The recordings were made in a small, sound-treated, anechoic room at the LDC. The speakers wore two microphones during the recordings, the first a Sennheiser HMD 410 headset, the second a Nortel Liberator wireless phone headset. The former's signal went through a Symetrix 302 Dual Microphone Preamp, Sony PCM-R300 DAT deck, and Townshend Datlink, then to a Sun Sparcserver 20 were it was written to (network) disk as 16 KHz, 16-bit, pcm data. The latter's signal, transmitted to a wireless base station at a telephone, connected via the telephone network to LDC's telephone recording platforms, where the digital data was captured to disk, ie. 8 KHz, 8-bit, u-law data.

The speakers were prompted via a computer interface which displayed one prompt at a time, and allowed the speaker to iterate through the prompts by pressing a "next" button. The task for each speaker was to read and say all their prompts, however this task was divided into multiple recording sessions, as the total task would take 2-4 hours, depending on the speaker's speech rate, error rate, etc. More specifically, the task was to pronounce all their prompts correctly (where correctly was a matter of judgement); prompt recordings deemed incorrect, or otherwise problematic, were re-recorded. These "redo" sessions were a combination of normal sessions, and "assisted" sessions, where a facilitator sat in the sound booth with the speaker to guide their pronunciation. A majority of the recording sessions were 15 minutes long, and the prompting program timed out after that long. Speakers sometimes did more than one session in a day, or more than one session in a row, depending on their availability, and how tedious they found the task. Initially, some sessions were done for 30 or even 60 minutes, but the session time was quickly reduced to 15 minutes due to the fact that generally people did become tired after even 15 minutes.

1.5 Presentation of Syllables via Prompts

It was deemed sufficient to collect a single token of each particular syllable a speaker was to say. However, due to differing research goals that end users might have, a "single token" was defined to be a phrase containing the syllable, plus the same syllable spoken in isolation. Therefore each prompt was of the form "I say blah now, blah" where blah represents the nonsense syllable. The prompting program created these prompts on the fly: it chose the syllable from a predetermined randomized list, unique to each speaker, and it chose three carrier words, inserting them all into the template given above. The template corresponded to, more or less, a "subject verb SYL adverb, SYL" sequence, where the first syllable position

Example	Darpabet	LDCbet	IPAbet
/a/ as in hat	@	xq	æ
/u/ as in hut	A	xa	ʌ
/ch/ as in church	C	xc	č / tʃ
/th/ as in this	D	xd	ð
/e/ as in bet	E	xe	ɛ
/ng/ as in sing	G	xg	ŋ
/i/ as in hit	I	xi	ɪ
/j/ as in judge	J	xj	ǰ / dʒ
/oy/ as in boy	O	xo	ɔɪ
/ir/ as in bird	R	xr	ɝ / əː
/sh/ as in she	S	xs	š / ʃ
/th/ as in think	T	xt	θ
/u/ as in put	U	xu	ʊ
/ow/ as in how	W	xw	aʊ
/y/ as in why	Y	xy	aɪ
/s/ as in pleasure	Z	xz	ž / ʒ
/o/ as in cot	a	a	ɑ / ɑː
/b/ as in bee	b	b	b
/au/ as in caught	c	c	ɔ / ɔː
/d/ as in dog	d	d	d
/ai/ as in bait	e	e	eɪ
/f/ as in fish	f	f	f
/g/ as in dog	g	g	g
/h/ as in he	h	h	h
/ee/ as in bee	i	i	i / iː
/c/ as in cat	k	k	k
/l/ as in look	l	l	l
/m/ as in man	m	m	m
/n/ as in man	n	n	n
/oa/ as in boat	o	o	o / oʊ
/p/ as in pen	p	p	p
/r/ as in real	r	r	r
/s/ as in see	s	s	s
/t/ as in cat	t	t	t
/oo/ as in boo	u	u	u /uː
/v/ as in vow	v	v	v
/w/ as in win	w	w	w
/y/ as in you	y	y	y / j
/z/ as in zoo	z	z	z

Table 2: *The above table provides the IPA symbol for each of the 1 and 2 character strings used to generate the names of the LDC sound files.*

could be considred an "object" position. The subject, verb, and adverb words were chosen randomly from sets of words appropriate for each position. This method created prompts with enough syntactic and semantic coherency to allow the speaker to fluently pronounce them. The speakers were instructed to say the phrase fluently, but to pause at the comma so that the second occurrence was truly isolated. Generally, the latter rule was enforced, but the former was not (see below).

The file "carrier.txt" contains the sets of carrier words that were used. These words don't precisely represent the actual prompts used, due to changes made along the way, but it's very close. The directory "prompts" contains one file per speaker, each file containing the actual prompts used for that speaker. This list is over 99% accurate, however there are the occasional cases where the prompt doesn't represent the actual words spoken for that syllable. In these files, each line contains the syllable, followed by the "|" character, followed by the actual text of the prompt. The syllables in the prompts were represented with real English words when they were also words; when they weren't, they were often represented with words that had parenthesized letters, which were interpreted as silent. Often various special representations were used to help elicit the correct syllable. The meaning of these representations should be obvious to the user of the corpus, when looking at the prompts, given the intended syllable, and are not worth making explicit here.

1.6 Auditing and Segmentation

Auditing and manual segmentation were performed on the recordings. In what follows, I use the word "prompt" to mean the recorded data elicited by some prompt, not the prompt itself. Timestamps were used to mark the beginning and end of each prompt, as well as to separate the phrase from the isolated syllable. These timestamps were then used to divide the audio data into individual files that contain either a single phrase or a single isolated syllable. A decision was also made as to the validity of each prompt, with invalid ones being added to the end of a speaker's list for re-recording (that is, the syllable from an invalid prompt was added, not the entire prompt). Mispronunciation of either instance of the syllable made the prompt invalid. However, exceptions were made if the auditor considered the "mispronunciation" to simply be the result of co-articulation and phonological variation, for example, a change in voicing due to an adjacent segment. Dialect variation was also permitted, please see the discussion on this below. There was no strict rule here, only a rule of thumb that it should sound like the intended syllable was pronounced.

Generally, not pausing at the comma made the prompt invalid, since the second instance of the syllable wasn't truly isolated. However, generally, non-fluent pronunciation of the phrase was allowed, meaning that often pauses were allowed within the phrase. No stance was held on co-articulation and its possible effect on the first instance of the syllable, in that this sort of effect (or lack thereof due to pausing) was not controlled for. However, to attempt to avoid any particular bias in the data in this respect, the use of a variety of words in the contextual word sets was intended to provide a variety of phonetic contexts, which of course were chosen at random.

The auditing, and therefore the segmentation, was performed on the wide-band recordings. The narrow-band recordings were not synchronized with the wide-band recordings, so to segment the narrow-band recordings using the same timestamps created for the wide-band recordings, the two recordings were aligned via a cross-correlation program.

1.7 Syllable Inventories

The files in inv/ represent the distribution (inventory) of syllables in the corpus by speaker and data type (wide-band vs. narrow-band). doc/inv_doc.txt documents the formats of these inventory files. There are 400 so-called "unique" syllables, 20 per speaker, and 2005 so-called "common" syllables, although in actuality only 1845 syllables are present for all 20 speakers. The remaining 160 syllables are missing from 1 to 3 speakers; that is, all common syllables are present for at least 17 speakers. These statements specifically describe the wide-band data. About 90% of the syllables recorded in the wide-band data were successfully recorded in the narrow-band data, but the other 10% were lost. The reasons for this lie in the particulars of the project design (ie. particular flaws), and to some extent the time constraints of the project. The above information, including the gaps in the narrow-band data, is represented in the files in inv/.

It should be noted that the list that each speaker began with was a combination of the common list and the speakers unique list, which was then randomized for that speaker. I say "began", because as mistakes were made, those syllables were added to the end of the list. The randomization of the lists should have minimized any session effects that may have been present, in the sense that the session effects wouldn't correlate with any other characteristic of the data. Furthermore, if one wanted to test for such session effects, they could not, since the ordering information is not reconstructable given the data on this disc. The ordering is however reconstructable from LDC internal data, if by chance someone was really interested in this information.

1.8 Filenames

doc/darbabet.txt contains the subset of the darpabet used to represent the syllables in data files. However, syllables were represented in a modified fashion in filenames: first, @ was replaced with Q, then all uppercase letters were replaced with a two character sequence, where the first character was "x" and the second character was the lowercase equivalent of the original character. The data is separated into two directories, "wb" and "nb", for wide-band and narrow-band recordings. These directories are subdivided for "phrasal" and "syllable (isolated)" cases, which are in turn subdivided by speaker. The phrasal vs. syllable distinction, shortened to "p" vs. "s", and the speaker distinction, as indicated by four character speaker IDs, are both represented in the filename, as well as the path. For example,

syls/wb/p/f101/p_f101_hxqt.sph syls/nb/s/f101/s_f101_xcxrxc.sph

represent two files by the same speaker f101, the first being a wide-band recording of a phrase containing the syllable "h@t" (hat), and the second being a narrow-band recording of an isolated occurrence of the syllable "CRC" (church). The gender of the speaker, "f" or "m", is encoded in the first character of their ID; the following digits uniquely identify the speaker, with the first digit always being 1, based on the idea that a second version of this corpus should use 2. The purpose of the syllable representation scheme was to maintain the syllable as part of the filename, such that the filename was file system safe, and that the syllable was still relatively readable by both human and machine. I think (and hope) the user will find this representation convenient, once familiar with the darpabet.

1.9 Issues of Dialect Variation

When possible, if it was recognized that someone had a particular dialect variation, their recordings were accepted if they pronounced the syllable faithfully for their dialect. For example, one dialect variation that appears in this corpus is the merger of the two high back vowels before /l/, such that "pull" and "pool" are homophonous (both sounding like the standard version of "pool", in my experience). So, for such a speaker, their pronunciations of "pull", which is [pUl] in standard speech but [pul] in their speech, won't be useful for those interested in syllables for their specific phonetic qualities. In other words, you can't use such a speaker's "pull" if you want the phonetic sequence [pUl]. Arguably, the speaker is not capable of producing such a sequence. However, for more abstract purposes, like the speech recognition of the word "pull", you can(!) use such a person's token, since it's a correct pronunciation of that word. This was one of the reasons such phonetic "unfaithfulness" was allowed, so that speech recognition research might capitalize on this instantiation of variation.

However, one serious problem emerged due to dialect variation, specifically relating to the so-called cot/caught merger. This merger, wide spread in the US and Canada, involves the two low back vowels [a] and [c], represented respectively by "cot" and "caught". Because some people on the project do not natively have this distinction, and because for those of us who do have it, like myself, the two vowels are still somewhat confusable, error entered the design of the project in at least two ways. First, many prompts for syllables with these vowels actually represented the opposite vowel, because the wrong word was chosen for the prompt. Second, many recordings were probably audited incorrectly. For example, if the desired syllable was /kat/, but they were prompted with the word "caught", the wrong syllable would have been elicited. In such a scenario, there would have also been a fair chance of the syllable being marked correct, even though it was not. This of course only affected speakers who could make the distinction; speakers with the merger would always produce the same vowel for these cases, as described above for pool/pull.

In most cases the correct vowel probably was elicited, but there are certainly errors due to incorrect prompt choices, more so with the vowels [a] and [c] than with other mergers. This problem became evident late enough in the project that it wasn't practical to remedy it. For those interested, the prompts (in the "prompt" directory) will aid in the determination of which vowel was actually elicited. Note that demographic information is given in doc/speaker_inf.txt, which is documented in doc/speaker_doc.txt.

1.10 Conversational Data

A small amount of conversational data was collected in addition to the syllable data. The description of the data is inlucded here, however, the data itself is not included on this disc (for space reasons), and is available on a separate CD.

About 10 minutes of conversational data was collected per speaker, although not for every speaker. The speakers were seated in the sound booth two at a time, one speaker wearing the same microphone setup for syllable recording, and the other wearing a third, lavalier mic. Each *_wb.sph file is a two channel file, containing the wide-band recording of the first participant, and the sole recording (wide-band) of the second participant, at 16 KHz, 16-bit pcm. The corresponding *_nb.sph file has the narrow-band recording for the first participant, at 8 KHz, 8-bit ulaw. This is the same scenario as the syllable recordings, except that the second participant was added as a second channel to the wide-band recording. The speakers

were given a list of possible topics, but were not forced to pick one of them, or any particular topic at all. The file topics.txt gives a rough idea of the topics chosen. Generally, two participants were recorded for five minutes, then asked to switch mics and pick a new topic, then were recorded for another five minutes. The filenames for the conversations represent the two speakers involved, as well as wide-band vs. narrow-band recording, ie. "wb" vs. "nb". I've here considered the "first" speaker to be the one wearing the microphones used for the syllable recordings. This speaker always corresponds to the first ID in the filename. I point this out because, for some reason, the first speaker doesn't always appear in the first channel in the audio data. In the case of the conversations, the narrow-band recordings were aligned with the wide-band recordings visually, using xwaves, rather than with the cross-correlation program, so are accordingly not as precisely aligned.

1.11 Acknowledgments and Contact Info

Many people deserve credit for the creation of this corpus. Mark Liberman, Jont Allen, Nelson Morgan, George Doddington, and others in the Novel Approaches group provided important conceptual advice in the design of the project. Chris Cieri provided continual guidance on many aspects of the project, both conceptual and practical. Dave Graff and Kevin Walker provided invaluable technical support. Special thanks to Dave Graff for assistance with the corpus documentation; much of the readability and accessibility of the documentation is due to him. Many of my fellow Penn Linguistics students provided crucial assistance with the work of the project, especially James Mesbur. As the implementor of this corpus, I take sole credit for all mistakes and shortcomings. Please don't hestiate to contact me with questions or problems, especially since we hope this corpus will prove to be the pilot for a subsequent, larger corpus. Your use of this corpus will allow us to determine the usefulness of data like this, and how such data might be improved. For all those who have been waiting for this corpus, thanks so much for your patience.

Jonathan Wright jdwright@ldc.upenn.edu 11/20/03