The Buckeye Corpus of Speech: Updates and Enhancements

Eric Fosler-Lussier[1,2], *Laura Dilley*[3], *Na'im Tyson*[2], *Mark Pitt*[4,2]

Departments of Computer Science and Engineering[1], Linguistics[2], and Psychology[4],
The Ohio State University, Columbus, OH, USA
Departments of Psychology and Communication Disorders[3],
Bowling Green State University, Bowling Green, OH, USA

fosler@cse.ohio-state.edu, dilley@bgnet.bgsu.edu, ntyson@ling.osu.edu, pitt.2@osu.edu

Abstract

This paper describes recent progress in the development of the Buckeye Corpus of Speech, a phonetically labeled corpus of conversational American English speech, first described in [1]. With the publication of the second phase of transcription, the corpus has nearly doubled in size from the first release. We briefly give an overview of the corpus, report on additional studies of inter-labeler agreement, and describe a new GUI designed to facilitate searching the annotated speech files.

Index Terms: corpora, transcription, phonetics, search tool

1. Introduction

The Buckeye Corpus of Speech was created to serve as a tool for expanding collective understanding of the problems that must be solved in order for humans and machines to recognize spoken words. Research over the past several decades using carefully articulated speech has been enormously useful for describing properties of spoken language and identifying key problems that must be answered. Study of informal speech should enrich these efforts, because it provides a more complete picture of the acoustics and phonetics of speech typically encountered by listeners.

The Buckeye Corpus is currently the largest extant phonetically labeled corpus of conversational American English. Approximately 300,000 words were collected from 40 talkers from the Columbus, OH area. Talkers in the corpus are counterbalanced for the gender of the talker (half are women, and half are men), for the age of the talker (half are under 30 years old, and half are over 40 years old), and for the gender of the interviewer. The speech is in interview format; talkers give monologues about various topics (the school system, politics, family life, etc.) in response to prompts from an interviewer. Interviews are approximately one hour in length; this is much more speech for an individual speaker than has been transcribed in projects such as the Switchboard Transcription Project [2], this focus on individual speakers allows for better understanding of *intra*-speaker pronunciation variability, and complements previous efforts that focused on *inter*-speaker variability [2].

Conversations were digitally recorded using a high-quality head-mounted microphone in a quiet room, and digital speech files have been recorded in uncompressed WAV format. The audio for each talker's conversation has been divided into a set of smaller audio files to facilitate manageable analyses and transcribed both orthographically and phonetically. The phonetic transcriptions were created in two stages. In the first stage, phonetic content was automatically aligned using the Xwaves Aligner program. In the second stage, trained phonetic analysts hand-corrected the automatically-generated phoneme alignments on the basis of spectrogram and waveform displays, as well as auditory perceptual information. The protocol for phonetic labeling was adapted from the TIMIT labeling guidelines [3]. Additional information about the speech collection and labeling can be found in [1] and in the Buckeye Corpus labeling manual [4].

	Phase 1	Phase 1+2
Number of speakers	20	40
Hours of interviews	19.7	38.1
Number of transcribed words	150915	296663
Number of transcribed phones	447788	870224

Table 1: Progress on the Buckeye Corpus as a function of number of speakers, hours of speech, number of words, and number of phones in two phases of public releases.

Transcription tiers are written in Xwaves format, so that they can be read with standard speech analysis software, e.g., Wavesurfer [5], Praat [6], and Xwaves (formerly from Entropics Inc.). In addition to word-level and phone-level alignments, log files contain miscellaneous information and notes about the speech or transcription process. ASCII text files that contain the orthography for all words plus labels for nonspeech sounds (e.g., markers for laughter, coughs, etc.) are also provided.

Due to the nature of the interviews, personal references often cropped up within the conversations. The corpus has been redacted to eliminate identifying references of the interviewees; this was achieved by "bleeping" the audio and removing the reference from the transcript.

The first phase of the project, described in [1], resulted in the transcription of speech from 20 speakers, with roughly an hour per speaker. In the remainder of this paper, we focus on the second phase of transcription, including inter-labeler reliability studies, as well as a search tool (SpeechSearcher) that facilitates working with such a large corpus.

2. Transcription, Phase 2

The second phase of transcription roughly doubled the amount of transcribed data (Table 1). In this second phase, fourteen undergraduate phoneticians were trained to transcribe the data using a coding scheme that was slightly revised from Phase 1 (see below for details). Otherwise, the transcription methodology followed our previously published guidelines [4]. Many of the student phoneticians were new to the project, so investigating the level of interlabeler agreement was a crucial component of the process.

Phoneme class	Inter-transcriber reliability test			
	(1) Pitt et al. ’05	(2) May ’05	(3) May ’06	Average
Overall	80.3%	78.5%	83.4%	80.8%
Vowels	73.6%	72.8%	87.1%	77.8%
Consonants	N/A	79.4%	80.9%	80.1%
Stops	92.9%	74.9%	85.3%	84.3%
Fricatives	91.2%	85.4%	78.0%	84.9%
Nasals	87.5%	77.1%	80.8%	81.8%
Liquids/Glides	86.5%	79.0%	74.7%	80.1%

Table 2: Agreement by phoneme class as measured in three inter-transcriber agreement studies. Results in column 1 correspond to the agreement reported in [1], while columns 2 and 3 correspond to agreement in two subsequent studies of inter-transcriber reliability. Results from the May 2005 study were used to retrain the phoneticians, resulting in generally better agreement in May 2006.

To ensure high consistency in phonetic alignment across the entire corpus, three tests of inter-labeler consistency were performed at several different stages during hand-correction of automatically-generated phoneme alignments. Detailed results of the first of these tests can be found in [1], while two subsequent inter-labeler consistency tests were conducted in May 2005 and May 2006. For the first test, a total of four phonetic analysts each transcribed four minutes of speech from the corpus; for the second and third tests, a total of eight and seven analysts, respectively, each transcribed two minutes of speech drawn from the corpus. The results of all three tests showed high reliability across major phoneme classes (see Table 2).

In all tests, labeling agreement was measured by calculating the proportion of pairs of labelers who assigned the same label to a given portion of speech, relative to all pairs of labelers. The tests provided a means of assessing overall cross-labeler reliability in use of the phonetic label set. Moreover, the first test [1] was used to determine whether certain phones in the phone set were used with relatively greater or lesser reliability than other phones. This test revealed that two sorts of vowel distinctions were difficult to identify reliably: (1) the reduced vowels [ə], [ɨ], and [ʉ]; and (2) the vowels [ə] and [ʌ], which differ in phonological status as reduced vs. full vowels but do not differ in vowel quality. The phone set was subsequently modified to collapse labels for [ə], [ɨ], [ʉ], and [ʌ] to a single ASCII label [ah] (for [ə]), and the corpus was standardized using this smaller phone set.

Subsequent tests of inter-transcriber agreement were used as practical metrics of quality assurance for phonetic labels. It can be observed from Table 1 that the results of all tests show good reliability across phoneme classes, with average agreement for almost all phoneme classes greater than 80%. This indicates high reliability consistent with previous findings of good inter-rater agreement (e.g., [7]). The slightly lower agreement in the second test was obtained at a point when new phonetic analysts had recently been hired onto the project. The data were used to provide direct feedback to these analysts on how to achieve greater overall agreement with other analysts, and subsequent labeling efforts in Summer 2005 focused on reanalysis of phonetic labels in order to achieve even higher agreement levels. The success of these efforts can be seen in the overall high agreement seen across phoneme classes in the third test and in averages across all three tests. Across the three tests, stops and fricatives showed the highest average agreement (84.3% and 84.9%, respectively), while vowels showed somewhat lower average agreement (77.8%). This may be related to the fact that consonants, especially stops, are perceived more categorically than vowels (e.g., [8, 9]).

3. SpeechSearcher: a corpus search tool

The first release of the Buckeye Corpus provided transcriptions in the standard XWaves format, which, as noted above, is readable by a variety of standard annotation tools. However, with a corpus this large, the capability of searching relatively quickly over multiple files of long (ca 30-60 minute) duration was sorely needed. The SpeechSearcher graphical user interface gives users the ability to find and manipulate sets of instances of word or phone sequences quickly within the corpus.

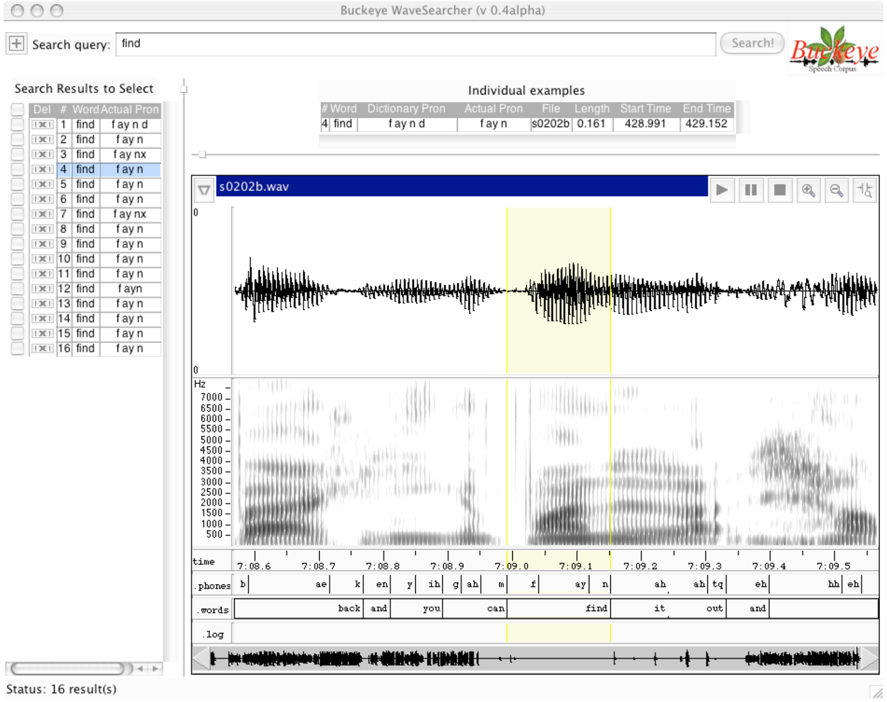
3.1. Interface Development

In developing the interface, we had four main desiderata. First, the interface must make it easy to find and display a large number of query results while enabling browsing of individual result instances within their original file context. Second, the interface should be familiar to many users, so that the learning curve for browsing results is not steep. Third, the program and database indices (although not the corpus data itself) must be packaged within a single application to facilitate installation and use and not require the downloading of dependent modules (such as a separate database engine). Finally, the software should be available across the Windows, Mac, and Linux platforms.

To address these issues, we decided to integrate a SQL database of indices over the Buckeye Corpus transcriptions with the WaveSurfer software package from KTH [5]. WaveSurfer is written primarily in Tcl/Tk, which makes it possible to wrap a search tool application (also written in Tcl/Tk) around the WaveSurfer windows; this means that we could utilize that familiar, open source package as our browser window. Much of the WaveSurfer functionality is retained within SpeechSearcher. Because the Buckeye Speech Corpus uses such large files, we extended WaveSurfer in two ways to speed up the interface. First, we modified the label-drawing routines so that the labels for the current window are drawn immediately upon displaying; the off-screen labels are drawn incrementally only when the application processing is idle. This significantly sped up loading of the files, particularly on slower machines. Second, we introduced caching of WaveSurfer browser windows to accelerate reloading of files previously visited within the session.[1]

The SQL database was implemented using SQLite [10], an embeddable SQL engine that is commonly included with many Tcl installations. The engine is relatively fast while not requiring users to install a separate database server on their machines.

[1] We also explored the idea of pre-caching WaveSurfer windows before they are displayed for the first time. Unfortunately the current version of the software does not seem to support drawing of windows off-screen and then displaying the final result.



Buckeye WaveSearcher (v 0.4alpha)

Search query: find

Search Results to Select

Del	#	Word	Actual Pron
☒	1	find	f ay n d
☒	2	find	f ay n
☒	3	find	f ay nx
☒	4	find	f ay n
☒	5	find	f ay n
☒	6	find	f ay n
☒	7	find	f ay nx
☒	8	find	f ay n
☒	9	find	f ay n
☒	10	find	f ay n
☒	11	find	f ay n
☒	12	find	f ayn
☒	13	find	f ay n
☒	14	find	f ay n
☒	15	find	f ay n
☒	16	find	f ay n

Individual examples

#	Word	Dictionary Pron	Actual Pron	File	Length	Start Time	End Time
4	find	f ay n d	f ay n	s0202b	0.161	428.991	429.152

Status: 16 result(s)

Figure 1: Example query for the word “find”, with sixteen search results (left column), displaying result number 4 in a Wavesurfer window (right column).

The corpus transcripts at the word, dictionary phone, and transcribed phone level were converted into SQL tables, with appropriate indices interrelating the labels as well as other information (e.g., speaker gender). Because joins of multiple phone sequences proved to be quite expensive (as there were many instances of each phone type), we also indexed multiple-phone sequences in the database to improve search times. Since there are many fewer instances of word types than phone types, indexing over multi-word sequences was not necessary.

3.2. Interface capabilities

When SpeechSearcher is first started, the user is provided with the *simple query* interface, which allows for searching over word sequences, dictionary phone sequences, or transcribed phone sequences. An example query (“find”) is shown in Figure 1. Successful queries provide a *result set* on the left side of the screen, which gives brief descriptions of the found word or word sequence instances (including the instance number, word sequence, and transcribed pronunciation). General operations on result sets are described below; clicking on any line n the result set displays the instance (all tiers plus visual displays) centered in its file context within the WaveSurfer window, as well as additional information (e.g., instance duration and file ID).

SpeechSearcher also supports an *advanced query* interface, where the user may search by any combination of the sequences listed above. Furthermore, the interface allows searching by segment length, speaker variables such as gender, age, or ID, interviewer gender, and/or time ranges within a particular file. The left side of Figure 2 shows a query for the word “this” transcribed with the phone [eh] in the subset of older male speakers. Queries in both the simple and advanced modes are saved in the query history list (and can be redeployed by selection in a menu); queries may also be saved or loaded from a file.

Queries, as noted above, provide result sets for browsing. Instances within a result set can be viewed, *marked* for later processing, or deleted from the result set. Result sets can be saved to or loaded from a file. In addition, marked results can be *exported* as individual segments (with corresponding phones and words files) for processing by another program. This makes SpeechSearcher an excellent tool for extracting examples from the corpus for linguistic studies.

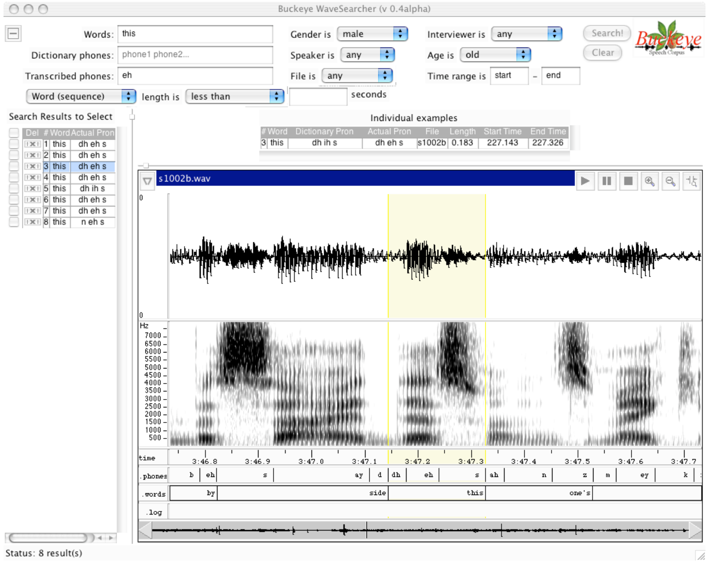
Finally, the program allows for viewing the current query as an SQL statement that is sent to the transcript database (right side of Figure 2). This allows advanced users to get some insight about the construction of the SQL database tables, and, with a bit of code examination, facilitates the construction of new queries within Tcl scripts for offline processing.

In future versions of the program, we plan to provide the ability to directly modify the relevant parts of the SQL query (so that advanced queried can be constructed). Other planned improvements are the ability to directly specify phonological rules to search for, as well as the integration of user-definable phonological feature classes.

In addition to being a resource for researchers, we hope that this tool will have pedagogical value in introductory courses for phonetics and speech technology. We welcome any feedback from the community on how this software and database are deployed in educational settings, and how we might improve their utility in the classroom setting.

4. Release information

The corpus and software packages can be downloaded from http://buckeyecorpus.osu.edu. The first release (March 2006)



Buckeye WaveSearcher (v 0.4alpha)

Words: this

Dictionary phones: phone1 phone2...

Transcribed phones: eh

Word (sequence) length is less than [ ] seconds

Gender is male

Speaker is any

File is any

Interviewer is any

Age is old

Time range is start – end

Search Results to Select

Del	#	Word	Actual Pron
☒	1	this	dh eh s
☒	2	this	dh eh s
☒	3	this	dh eh s
☒	4	this	dh eh s
☒	5	this	dh ih s
☒	6	this	dh eh s
☒	7	this	dh eh s
☒	8	this	n eh s

Individual examples

#	Word	Dictionary Pron	Actual Pron	File	Length	Start Time	End Time
3	this	dh ih s	dh eh s	s1002b	0.183	227.143	227.326

s1002b.wav

Status: 8 result(s)

```
SELECT words_main.word              AS 'word'
  ,words_main.segment_id            AS 'segment'
  ,words_main.set_id                AS 'set'
  ,sets_main.speaker                AS 'speaker'
  ,sets_main.directory              AS 'directory'
  ,segments_main.filename           AS 'filename'
  ,segments_main.length             AS 'segment_length'
  ,words_main.time_start            AS 'time_start'
  ,words_main.time_end              AS 'time_end'
  ,words_main.length                AS 'length'
  ,words_main.id                    AS 'order'
  ,words_main.phone_count           AS 'phone_count'
  ,words_main.dict_phone_count      AS 'dict_count'
  ,words_main.worddict              AS 'worddict'
  ,words_main.wordphone             AS 'wordphone'
FROM words                          AS words_main
  , sets                            AS sets_main
  , segments                        AS segments_main
  , phones                          AS phones_main
WHERE words_main.word = "this"
  AND phones_main.phone = "eh"
  AND phones_main.word_id = words_main.id
  AND phones_main.segment_id = words_main.segment_id
  AND sets_main.gend = "m"
  AND sets_main.age = "o"
  AND sets_main.id = words_main.set_id
  AND segments_main.id = words_main.segment_id
```

Figure 2: (left) Advanced search interface, showing database entries of the word “this” pronounced with the phone [eh], restricted to older male speakers. (right) The resulting SQL query as displayed by the interface.

[11] contained the first half of the corpus (20 talkers) while the second half of the corpus (remaining 20 talkers) was released in February 2007 [12]. The SpeechSearcher software is due to be released in April 2007. The corpus and software are available to researchers, free of charge, via direct download or DVD exchange. For more details, see the Registration page at buckeyecorpus.osu.edu.

5. Acknowledgements

We thank the many students who served as labelers. In addition, thanks go to Joe Ryan, who provided an initial prototype of the SQL database, Jessica Fosler and Peter Coleman for comments on the user interface, and Kåre Sjölander and Jonas Beskow for providing WaveSurfer to the community as an open-source resource. This work was supported by research grant DC004330 from the National Institute on Deafness and Other Communication Disorders.

6. References

[1] M. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, “The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability,” *Speech Communication*, vol. 45, pp. 90–95, 2005.

[2] S. Greenberg, J. Hollenbach, and D. Ellis, “Insights into spoken language gleaned from phonetic transcription of the Switchboard Corpus,” in *International Conference on Speech and Langauge Processing*, Philadelphia, PA, 1996.

[3] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, “DARPA TIMIT acoustic-phonetic continuous speech corpus,” Tech. Rep. NISTIR 4930, NIST, Gaithersburg, MD, 1993.

[4] S. Kiesling, L. Dilley, and W. Raymond, *The Variation in Conversation (ViC) Project: Creation of the Buckeye Corpus of Conversational Speech*, Department of Psychology, Ohio State University, Columbus, OH, 2006.

[5] K. Sjölander and J. Beskow, “Wavesurfer – an open source speech tool,” in *Proceedings of ICSLP*, Beijing, 2000.

[6] P. Boersma and D. Weenink, *Praat, a system for doing phonetics by computer*, http://www.praat.org, 2002.

[7] R. B. Irwin, “Consistency of judgments of articulatory productions,” *Journal of Speech and Hearing Research*, vol. 13, pp. 548—555, 1970.

[8] A. F. Healy and B. H. Repp, “Context independence and phonetic mediation in categorical perception,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, no. 1, pp. 68–80, 1982.

[9] M. E. H. Schouten and A. J. van Hessen, “Modeling phoneme perception. I: Categorical perception,” *Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 1841–1855, 1992.

[10] “Sqlite software library,” http://www.sqlite.org, March 2007.

[11] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, “Buckeye Corpus of Conversational Speech (2006; 1st release),” http://www.buckeyecorpus.osu.edu, 2006.

[12] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, “Buckeye Corpus of Conversational Speech (2007; 2nd release),” http://www.buckeyecorpus.osu.edu, 2007.