Low-cost Customized Speech Corpus Creation for Speech Technology Applications

Kazuaki Maeda, Christopher Cieri, Kevin Walker

Linguistic Data Consortium
University of Pennsylvania
3600 Market St., Suite 810
Philadelphia, 19104 PA, U.S.A.
{maeda, ccieri, walkerk}@ldc.upenn.edu

Abstract

Speech technology applications, such as speech recognition, speech synthesis, and speech dialog systems, often require corpora based on highly customized specifications. Existing corpora available to the community, such as TIMIT and other corpora distributed by LDC and ELDA, do not always meet the requirements of such applications. In such cases, the developers need to create their own corpora. The creation of a highly customized speech corpus, however, could be a very expensive and time-consuming task, especially for small organizations. It requires multidisciplinary expertise in linguistics, management and engineering as it involves subtasks such as the corpus design, human subject recruitment, recording, quality assurance, and in some cases, segmentation, transcription and annotation. This paper describes LDC's recent involvement in the creation of a low-cost yet highly-customized speech corpus for a commercial organization under a novel data creation and licensing model, which benefits both the particular data requester and the general linguistic data user community.

1. Introduction

The Linguistic Data Consortium (LDC) is a non-profit organization, whose primary mission is to support education, research and technology development in language-related disciplines. LDC creates and disseminates linguistic resources for this mission; it does not create linguistics resources to benefit a single organization. There have been, however, strong interests from commercial and non-commercial organizations to subcontract LDC to create customized speech corpora. In response to such data requests from organizations while meeting its mission goals, LDC has created a ”delayed release” model of data creation and licensing. Under this model, the data requester (also the sponsor) subcontracts LDC to create a speech corpus meeting their specifications. The data requester funds the creation of the corpus, and in return, benefits from a lead time of typically eighteen months, in which the data requester has the exclusive rights to use the data. The corpus is customized to their needs, and the effort of communicating needs to an outside group generally has a clarifying effect; in the process, they may learn of approaches or technologies they had not considered. After the lead time, which begins when the corpus is delivered, LDC releases the corpus to LDC members and non-members at a significantly reduced cost.

2. Speech Controlled Computing Corpus

The Speech Controlled Computing (SCC) corpus was the first corpus created at LDC under this model. It was developed for limited-vocabulary speech recognition applications targeting a wide variety of American English speakers. The data requester's idea was to have a set of recordings of isolated words and short phrases of the kind one would use to control household appliances recorded to be representative of most American speakers. LDC and the data requester agreed to have a pool of speakers that represented each of four regional groups, three age groups and two gender groups. To meet this first challenge, we conducted a recruitment effort to meet these demographic requirements. In order to ensure that recordings are consistent and of highest possible quality, all recordings were to be done in a recording booth in the LDC suite. This limited the pool of possible participants to Philadelphia-region residents. Another challenge we faced was the paucity of male subjects willing to participate, particularly in the mid and high age groups. One of the most effective solutions to this problem was to recruit subjects who had participated in our previous studies, such as Fisher and Mixer. LDC keeps a database of speakers who participated in the past projects. LDC's human subject recruitment team contacted possible participants living in the Philadelphia region for each of the 24 demographic groups shown in Table 1.

Region	Male			Female		
	Young	Mid	Old	Young	Mid	Old
North	7	6	3	6	6	5
South	5	5	3	7	3	6
Midland	9	8	6	8	7	5
West	5	3	1	5	3	2

Table 1: Speakers in the SCC corpus

The recordings were performed in LDC's sound-attenuating recording room. Prior to the production recordings, we conducted a series of pilot recordings to test whether our recording method met the requirements set by the data requester. A number of microphones, including headset-mounted microphones and stand-mounted microphones, were tested, and sample recordings were sent to the data requester to determine the best recording method for them. In order to facilitate efficient recordings, LDC developed an infrastructure to control and monitor recording sessions, the hardware components of which include hard disk-based digital recording with back-up recordings to DAT tapes. A software-based prompter was created to display the word list in a randomized order to speakers, and to control the progress of the recordings. An assistant recording engineer monitored the recording sessions outside of the recording booth. Table 2 shows a partial list of the words recorded for the SCC corpus.

alarm	answer
answer	arm
back	balance
bass	brake
call	camera
cancel	CD
channel	clear
close	computer
control	cook

Table 2: A partial word list for SCC

The digitized sound files were initially segmented into individual tokens using an automatic acoustic segmentation program. As the corpus specifications required all tokens to be manually reviewed and for the segmentation to be corrected, we created a specialized annotation tool for these purposes. The resulting auditing and segmentation tool utilizes the Annotation Graph Toolkit (AGTK), Snack and WaveSurfer (Sjölander and Beskow, 2000). The WaveSurfer module provides the ability to play back audio, display waveforms and compute spectrograms. The spectrograms provide the auditor/segmenter an effective means to judge what is spoken and where each utterance begins and ends. Our experience demonstrates that LDC annotators, even those who are not necessarily trained phoneticians, were able to recognize the key features of spectrograms after a short training period. However, student workers with a strong interest in languages and/or linguistics were particularly well suited for this task.

3. Delayed Release Licensing Model

The delayed release model used for the SCC corpus benefits both the individual data requester and the international user community of linguistic resources. In this model the data requester, who is also the sponsor of the project, receives a lead time of typically 18 months. During this time, the data requester has the exclusive rights to use the data. After this time, the data set is published as an LDC general publication to LDC members and non-members. This is documented in the statement of work, and is agreed upon between the data requester and LDC at the time of the contract.

The data set released to the LDC members and non-members is essentially identical to what is delivered to the data requester. In the case of the SCC corpus, only the following changes were made to the general publication:

- File names used to identify the names of the subjects were changed to anonymous subject names.
- Mentions of sponsor names were removed from the documentation.
- The regions of silence before and after each utterance were extended from 10 ms to 100 ms.

Figure 1 illustrates the delayed release licensing model.

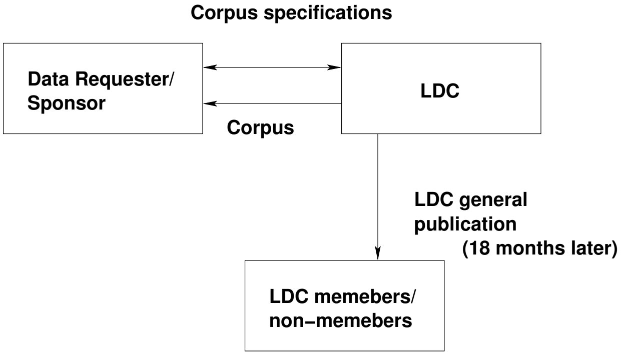

Corpus specifications
Data Requester/ Sponsor
LDC
Corpus
LDC general publication (18 months later)
LDC memebers/ non−memebers

Figure 1: Delayed Release Licensing Model

4. Carpooling Philosophy

The carpooling philosophy here refers to the grouping of potential data requesters/sponsors in order to reduce the cost of customized speech corpus creation. Many speech corpus creation efforts are similar in terms of speaker and recording requirements. For example, if three projects require 50 native speakers of English to be recorded in a sound booth, it is much more cost and time efficient for us to record each subject for all of these projects at the same time, than to have each speaker come in three times.

Similarly the corpus design for multiple projects may be very similar so that we can merge multiple data sets into one. This will also significantly reduce the cost for the data requesters.

5. Solidifying In-house Infrastructure and Knowledge Base

5.1. Overview

In order for this venture to be successful, LDC will need to solidify its expertise in customized speech corpus creation. LDC's existing infrastructure and knowledge bases will need constant enhancement and improvement. In the following sections, we discuss not only our current areas of expertise in creating unique speech corpora, but also our plans for strengthening our approaches and methods in this effort.

5.2. Corpus Design

The first step in creating a customized speech corpus is to design the corpus (Gibbon et al., 1997). This should be discussed throughly between the data requester and LDC. While the data requester may or may not be an expert in corpus design methodologies, LDC employs experts in various fields of Linguistics, including Sociolinguistics and Phonetics, as well as experts in Speech Technologies, and can assist the date requester to define the corpus needs.

5.3. Human Subject Recruitment

In the past, the LDC has recruited human subjects to participate in both telephone conversation recording sessions and in-house recording sessions. The current Mixer study records native speakers of various languages[1] (Cieri et al., 2006). The participants in these studies are likely to be interested in participating similar projects. We keep a database of our past participants and inquirers, and expect this database to grow. In addition, LDC, as part of the University of Pennsylvania, has access to student populations from various parts of the country and the world.

5.4. Recording Infrastructure and Methodologies

LDC offers an extensive telephone speech collection system, as well as facilities to record subjects in-house. In-house recordings are made in a sound attenuated recording booth. The recording booth is equipped with stand-mounted microphones and headset-mounted microphones as well as a monitor display, which may be used to show software-controlled prompts. The recorded sounds are sent to the monitoring station outside of the booth. The recording and digitization can be made to both a DAT tape and a hard disk drive. LDC has a multi-channel digital recording system that can be used to record multiple speakers simultaneously, or to record a single speaker with multiple microphones.

The current setup requires a recording assistant to monitor the recording level and the speaker's pronunciation. If a particular word or phrase needs to be repeated, the recording assistant operates the prompting software to show the prompt for the same word or phrase again. Some of these tasks, such as monitoring the recording level could be automated. We plan to create an automated real-time quality control system that checks for recording problems.

5.5. Auditing and Annotation

LDC has expertise in transcription and annotation of spoken corpora by maintaining a well-training staff of transcribers, annotators and experts in these fields. In addition, LDC has in-house software developers who are well experienced in designing and developing customized annotation tools (Maeda et al., 2006; Bird et al., 2002), such as the auditing and segmentation tool used in the creation of the SCC corpus. The tool displays a spectrogram which allows the auditor to identify words and word boundaries visually, as shown in Figure 2.

5.6. Financial Considerations

Recruitment, recording methodologies and annotation are only part of of the picture of speech corpus design. Additionally, LDC must estimate the cost required for a customized speech corpus creation effort. If the cost estimate is too low, LDC will lose money and time of the staff members involved. If the cost estimate is too high, then it may be difficult to attract subcontractors to LDC.

It is also important for LDC to compensate the subjects at the right rate. We consider these subjects to be an extremely valuable resource; we hope that the subjects will come back and participate in future studies. Both undercompensation and overcompensation hurt us in the long run.

The experience from the SCC corpus creation as well as from studies, such as Mixer and Fisher, gave us some good ideas about these aspects of speech corpus creation. We will analyze the financial aspects of each study at various stages, and will incorporate the results into our knowledge base.

6. Commercial Sponsorship and Other Possibilities

The delayed release licensing model allows LDC to create speech corpora for commercial organizations. This model may also be an attractive option for research and academic organizations who need to create speech corpora for their research. The lead time allows the researchers use the data exclusively and publish the results. The researchers then can cite the LDC data publication as the data used in their study, allowing the readers to access their data.

7. Conclusion

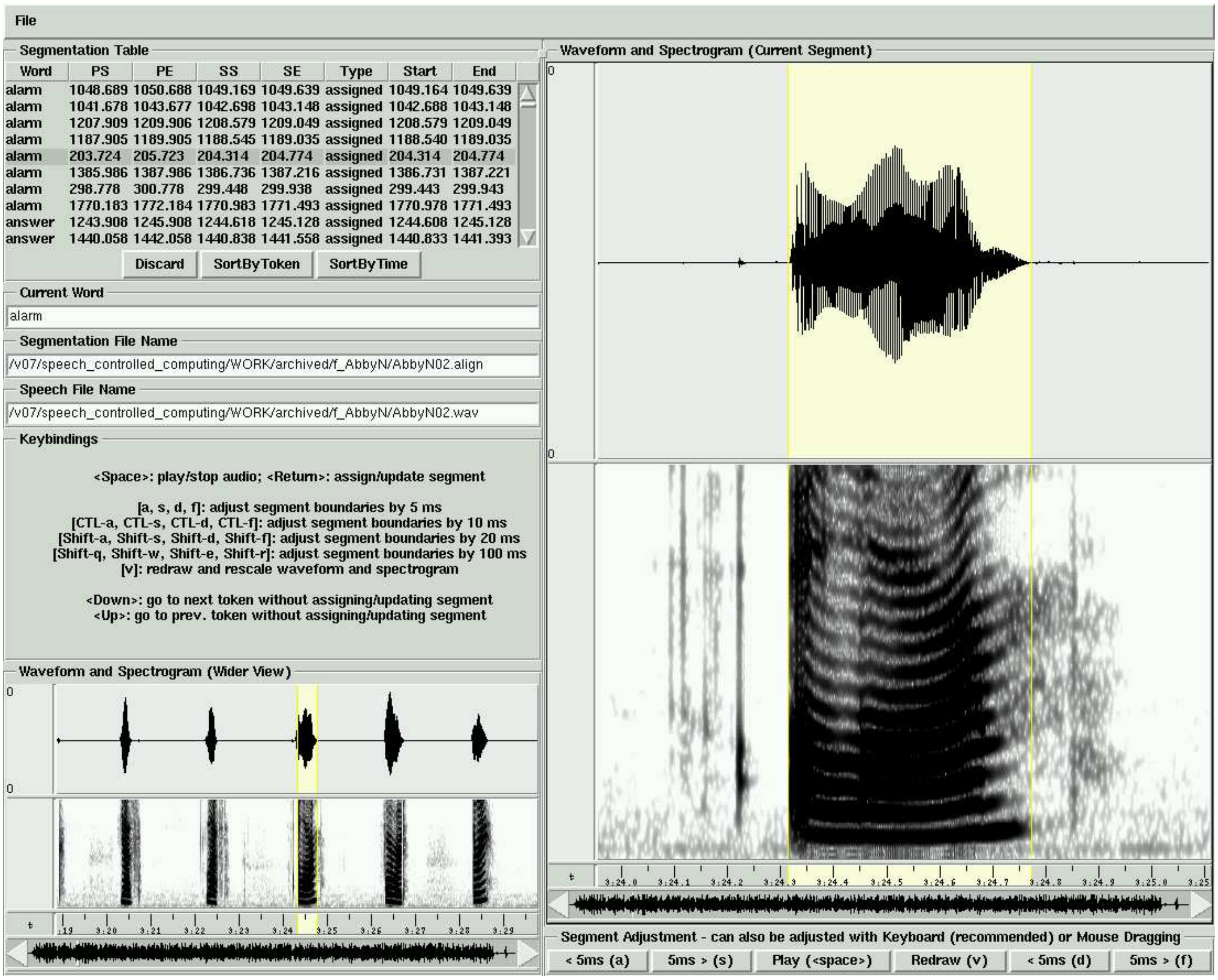
Speech corpus creation under the new delayed release licensing model presents a number of attractive advantages to the data requester, LDC and the linguistic data user community. The data requester indeed receives the customized corpus they require utilizing LDC's multidisciplinary expertise in linguistic data creation. LDC, on the other hand, seizes the opportunity to strengthen its expertise, and in some cases to enhance its infrastructure and to add to its subject database, all of which benefit its user community including future commercial users. The linguistic data user community, speech technology researchers and linguistic researchers alike, are able to to access the data through LDC publications after the lead time was passed. LDC has created speech corpora of various types, including telephone conversation recordings, meeting recordings, multimodal recordings and other specialized corpora, such as the Emotional Prosody speech corpus, in which actors and actresses simulated various emotional speech. We expect that the new data creation model provides LDC new opportunities to utilize our expertise in creating various speech corpora.

8. References

Steven Bird, Kazuaki Maeda, Xiaoyi Ma, Haejoong Lee, Beth Randall, and Salim Zayat. 2002. TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse tools built on the Annotation Graph Toolkit. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

Christopher Cieri, Walt Andrews, Joseph P. Campbell, George Doddington, Jack Godfrey, Shudong Huang, Mark Liberman, Alvin Martin, Hirotaka Nakasone, Mark Przybocki, and Kevin Walker. 2006. The Mixer and transcript reading corpora: Resources for multilingual, crosschannel speaker recognition research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

[1] http://mixer.ldc.upenn.edu



Word	PS	PE	SS	SE	Type	Start	End
alarm	1048.689	1050.688	1049.169	1049.639	assigned	1049.164	1049.639
alarm	1041.678	1043.677	1042.698	1043.148	assigned	1042.688	1043.148
alarm	1207.909	1209.906	1208.579	1209.049	assigned	1208.579	1209.049
alarm	1187.905	1189.905	1188.545	1189.035	assigned	1188.540	1189.035
alarm	203.724	205.723	204.314	204.774	assigned	204.314	204.774
alarm	1385.986	1387.986	1386.736	1387.216	assigned	1386.731	1387.221
alarm	298.778	300.778	299.448	299.938	assigned	299.443	299.943
alarm	1770.183	1772.184	1770.983	1771.493	assigned	1770.978	1771.493
answer	1243.908	1245.908	1244.618	1245.128	assigned	1244.608	1245.128
answer	1440.058	1442.058	1440.838	1441.558	assigned	1440.833	1441.393

Current Word: alarm

Segmentation File Name: /v07/speech_controlled_computing/WORK/archived/f_AbbyN/AbbyN02.align

Speech File Name: /v07/speech_controlled_computing/WORK/archived/f_AbbyN/AbbyN02.wav

Keybindings

<Space>: play/stop audio; <Return>: assign/update segment

[a, s, d, f]: adjust segment boundaries by 5 ms
[CTL-a, CTL-s, CTL-d, CTL-f]: adjust segment boundaries by 10 ms
[Shift-a, Shift-s, Shift-d, Shift-f]: adjust segment boundaries by 20 ms
[Shift-q, Shift-w, Shift-e, Shift-r]: adjust segment boundaries by 100 ms
[v]: redraw and rescale waveform and spectrogram

<Down>: go to next token without assigning/updating segment
<Up>: go to prev. token without assigning/updating segment

Figure 2: SCC Auditing and Segmenting Tool

Dafydd Gibbon, Roger Moore, and Richard Winski, editors. 1997. *Handbook of Standards and Resources for Spoken Language Systems: Spoken Language System and Corpus Design*, volume 1. Mouton de Gruyter.

Kazuaki Maeda, Haejoong Lee, Julie Medero, and Stephanie Strassel. 2006. A new phase in annotation tool development at the linguistic data consortium: The evolution of the annotation graph toolkit. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Kåre Sjölander and Jonas Beskow. 2000. WaveSurfer – an open source speech tool. In *Proceedings of the 6th International Conference on Spoken Language Processing*. http://www.speech.kth.se/wavesurfer/.