Research Methodologies, Observations and Outcomes in (Conversational) Speech Data Collection

Christopher Cieri
University of Pennsylvania
Linguistic Data Consortium
3615 Market Street
Philadelphia, PA 19026
+1 215-573-5489
ccieri@ldc.upenn.edu

David Miller
University of Pennsylvania
Linguistic Data Consortium
3615 Market Street
Philadelphia, PA 19026
+1 215-573-9182
damiller@ldc.upenn.edu

Kevin Walker
University of Pennsylvania
Linguistic Data Consortium
3615 Market Street
Philadelphia, PA 19026
+1 215-573-4172
walkerk@ldc.upenn.edu

ABSTRACT

This paper presents research methodologies for collecting speech data and gives observations from a recent set of conversational speech collections before describing their outcomes. The presentation begins with a comparison of the relative challenges offered by broadcast news, telephone conversation and meeting recordings. The remainder of the discussion focuses on methods for collection of conversational data with special focus on two recent Switchboard collections. We identify method that have allowed for very cost-efficient collection of Switchboard data. We conclude with a summary of generally available resources that result from the efforts described herein.

Keywords

Speech data, audio, telephone collection, meeting data, human subjects, ASR, collection methodology, speaker identification, speech recognition

1. Introduction

Large volumes of speech data are crucial to progress in speech technologies such as speech recognition and speaker identification, to research programs that rely on speech recognition output such as: Topic Detection and Tracking (Wayne 1998, 2000), Automatic Content Extraction (NIST 2000) and to applications such as speech-to-speech translation. This paper discusses methodologies for collecting speech data that have evolved over dozens of collection projects involving thousands of speakers. It begins by identifying challenges present in collecting digital audio from both broadcast news and conversational speech. The discussion focuses primarily on techniques for collecting conversational speech. Observations of human subjects' responses to collection procedures identify methods that yield more natural data while reducing cost. All of the data sets described herein are or will be available for use in linguistic education, research and technology development.

2. Communicative Interactions in Data Supporting Speech Technology Development

Recordings of human communicative interactions used in speech research vary along a number of dimensions. The acoustic properties of a speech signal vary with the channels through which the speech passes on the way from speaker to hearer. Each human language presents its own set of problem including new phoneme inventories and word formation rules and co-occurrence patterns to be modeled. Differences in communicative situation are reflected in vocabulary, speech rate and voice quality. Demographic factors such as age, sex and region of origin affect phonology, lexical choice and sentence formation. Finally the application in which the recognition technology will be used affects vocabulary size and the requirements for the system's speech, accuracy and output format. These concerns are equally present for recognition of broadcast news, telephone conversations and multi-party meetings.

There are other dimensions along which broadcast news, telephone conversations and multi-party meetings vary, rendering each either more or less challenging to speech researchers. Figure 1 lists several of the dimensions and indicates how they affect the degree of challenge each type presents to human annotators and ASR systems alike. Here we will describe just a subset. Rows one and two refer to the degree of variability present in the physical environment and in the audio capture equipment used. In broadcast news, most speech takes place in the studio environment where high quality equipment and quiet prevail. In contrast, recognition of telephone speech suffers from the greater variability present in the physical environment in which the speakers find themselves. Rows three and four point to both movement in place and change of location as factors affecting the difficulty of speech data. Broadcast news personalities tend to sit in a single place and minimize movements that would create noise. Conversational speech lacks this level of discipline. Not only may participants generate additional noise through movements but they may also change their location relative to the data capture devices by moving a phone away from their mouths, by walking out of the range of a wireless phone base or, in the meeting environment, by walking alternately toward and away from room microphones. Broadcast news does present a greater challenge than telephone conversation in its multi-modal signal. The modern television broadcast may contain not only the video of the on-air personality but also background images, closed captioning, sidebar text and the horizontally scrolling text, or "crawl", that CNN has conspicuously employed recently. Integrating these sources of information is an open research problem for information management technologies. Broadcast news speech, relatively formal and well-rehearsed, contains a narrower variety of linguistic styles and fewer disfluencies and rapid speech phenomena than conversational speech. In telephone conversations the number of speakers is usually small and fixed while in broadcast news there may be studio-guests, call-ins and man-on the-street interviews. The *information handicap* in Figure 1 refers to the paucity of information the annotator or recognition system has relative to the participant in a communicative interaction. For telephone conversations, a recognition system has as much signal data as the interlocutors. However in meetings and broadcast television, the facial expressions, maps, etc. that help to disambiguate the audio for participants are lacking in the audio signal. The *Observer's Paradox* states that in order to understand human communication one must study it but the very act of observation affects the phenomenon under study. One assumes that in broadcast news the effect of observation is essentially zero. In LDC telephone collections, there is both evidence that participants believe they should monitor their speech and evidence that they sometime forget to do so. The effect of observation has the potential to be the most profound in meetings where special rooms may be required and where microphones may be in plain site.

	Broadcast	**Telephone**	**Meetings**
Variable Environment		🟨	🟥
Variable Capture		🟨	🟥
Movement		🟨	🟥
Change of Location		🟨	🟥
Multimodality	🟥		🟥
Informality		🟥	🟥
Impromptu Speech		🟥	🟥
Overlapped Speech			🟥
External Apparatus	🟨		🟥
Multiple Speakers	🟨		🟥
Information Handicap	🟨		🟥
Observer's Paradox			🟥
Readable Transcript	🟨	🟨	🟥
Increasing Challenge =>		🟨	🟥

Figure 1: Comparison of Human Interactions underlying three speech data types.

3. Collection Types

In the sections that follow, due to time and space limitations, we will focus our attention on collections of conversational speech, specifically on Switchboard style telephone collections and GroupTalk and GroupMeet style collections of meetings.

The Switchboard call collection protocol is optimal for speaker identification research although researchers in speech recognition, discourse analysis and sociolinguistics, among others, have used Switchboard data. Switchboard targets from 200 to 700 speakers each participating in 10 telephone calls with other participants whom they, typically, do not know. Calls last 5-6 minutes and interlocutors discuss assigned topics. These features distinguish Switchboard from the CallHome and CallFriend protocols that are optimized for large vocabulary speech recognition. CallHome and Call Friend target 100 to 200 participants each of whom makes a single 20-30 minute call to a close friend or family member. Participants are native speakers of the target language of the study; topics are not constrained.

GroupTalk and GroupMeet are protocols for collecting conversational speech during face-to-face meetings. GroupTalk collects facilitated discussions. Under the GroupTalk protocol, an interviewer joins a group of 2 to 4 friends or family members introducing a variety of topics with the goal of identifying a few that truly interest the group and engage them in extended discussions. During GroupTalk sessions we tend to use less obtrusive microphones. Successful GroupTalk sessions will contain relaxed and informal speech. GroupMeet targets planning meetings. The goal of GroupMeet is to identify groups that meet regularly or were planning to hold a specific meeting and gain their permission to record that meeting for research purposes. During GroupMeet sessions, we deploy a variety of microphones. Speech tends to be more formal than in Group Talk.

4. Collection Procedures

A conversational data collection can be divided into two obvious phases, recruitment and collection. Participant recruitment is a crucial and sometimes overlooked aspect in this process for, without participants, there are no data. Our experience over multiple Switchboard collections has shown that several factors in the recruitment process can have profound effects on the collection's outcome. These include 1) determining the best time of year to undertake a collection 2) determining the hours of the day when collection is possible 3) the number of recruits necessary 4) setting the compensation to encourage participants to complete the terms of the study. During the first Switchboard Cellular (1999-2000), our requirement that participants make a certain number of phone calls while outdoors lead us to begin this study three months earlier than we had hoped simply to avoid beginning the study during the coldest months of winter. Under the Switchboard model, we also observed that restricting the hours during which participants can make calls raised the probability that they would actually reach another available participant.

Ideally, participant recruitment occurs a few weeks prior to a collection's commencement. If recruitment occurs too far in advance of a collection's commencement, participant interest is likely to wane. If a telephone collection begin too soon after recruitment begins, the lack of a critical mass of available participants may frustrate callers. The best recruitment efforts, however, are only as good as the technology that supports them. Before recruitment begins, it is imperative to have a reliable subject database and a user-friendly interface to support the recruitment team. Subject data includes: name, gender, age, education, country born/raised in. For purpose of payment and participant care, social security number and contact information are also crucial. Generally speaking, this data is collected during the initial discussion between the participant and the recruitment staff. Indeed, for a telephone collection, that may also be the only time the recruiters speak directly with a participant.

LDC generally advertises via print media and electronic announcements. Potential participants then contact the LDC via phone or e-mail whence they learn: 1) that speech will be recorded for research and educational purposes, 2) that personal information will be kept confidential, not be released with the data, 3) when the study begins and ends and how to participate, 4) how, how much and when they will be compensated.

After initial contact with the recruitment staff, each potential participant receives a set of written detailed instructions that reiterate everything above. In telephone studies, the instructions include the participant's unique number (PIN) and the series of prompts that will be heard. Certain conversational studies (Switchboard) require a critical mass of recruits before they can begin. In other studies, such as CallHome and CallFriend, a participant can begin immediately after registering. For GroupTalk and GroupMeet studies, where either subjects come to LDC or LDC staff go to the subjects' meeting room, the subjects sign an informed consent form. In telephone collections, participants indicate their consent to have their voices recorded multiple times. After receiving their written instructions, subjects must call the robot operator to indicate their consent and activate their PINs. When making or receiving a call subject must again indicate their consent to be recorded. In each type of study, participant compliance is closely monitored to ensure a successful study. If a study does not proceed according to plan, adjusting study parameters including the number of recruits, their demographics and their compensation may be helpful

5. Collection Technology

Conversational speech collection systems must be accurate, reliable, economical, and capable of delivering real world data. For both telephone speech and meeting collection, LDC has developed a robust system that leverages off-the-shelf hardware. The system consists of customized software, telephony hardware, and a project database and can record multiple simultaneous conversations with no need for operator intervention. The project database containing demographic information and call activity statistics for each participant, supports all recruitment, collection and reporting software. The demographic information is entered during recruitment. The call activity statistics are updated each time a participant tries to make a call, or receives one; the call logic software requires accurate, timely information.

The current collection platform is a Windows NT server with a high capacity RAID, Dialogic telephony hardware, and telephony libraries/API. Calls arrive via leased T-1 line. The platform can record up to 12 simultaneous conversations. Processes such as participant validation, T-1 signaling, initiation and termination of recording, and error handling are processed by call logic software developed at LDC. The telephony software is written in VOS, a programming language developed by Parity Software (this has since been purchased by Dialogic and re-branded as CT-ADE). VOS is easy to use, but provides the flexibility to develop complex and robust applications. Based on our experience so far, VOS is quite reliable.

In an effort to accommodate data collection informants, several improvements have been made to the LDC's telephony applications. In the case of the Switchboard application, we have tried to make the participation process less onerous by improving database performance, adding music on hold, and allowing participants to check their participation statistics when they call into the platform. We have also redesigned the callee selection routine. A caller on hold is notified as each potential callee is polled. After a certain number of callees have been tried, the caller is given an opportunity to quit and try their call at a later time, or continue waiting. Adopting a customer-service approach to participants encourages them to complete their role in the study.

The LDC's meeting recording system can record 16 tracks of digital audio. The system features a mixture of wireless and far-field wired microphones. Depending upon the session, either lavalier or head-mounted microphones are used for close-micing of each participant. Room microphones, including a microphone array, PZM, omni-directional and directional microphones are also used. The meeting recording system consists of a digital mixer, a multi-track digital tape recording deck, wireless microphone receivers, a microphone preamplifier, and a multi-channel digital audio computer interface. Meeting sessions are recorded as 16bit/44kHz PCM audio.

6. Observations

To date, LDC has conducted five Switchboard style collections in-house: Switchboard 2 Phases 1, 2 and 3 and Switchboard Cellular Phases 1 and 2. Switchboard 2 Phase 1 included 657 speakers most of whom were residents in the Mid-Atlantic area: (PA=303, NJ=116, NY=53, DE=13, CT=12, MD=14, OH=13, MA=8). Many of the participants in SWB-2 Phase I were college students from the following universities: Penn State University, University of Delaware, University of Pennsylvania, Drexel University, and Rutgers University. However, there was also a large contingent of local Philadelphians in Phase 1. Of the 657 participants, 358 were female and 299 were male. Switchboard 2 Phase 2 included 4,472 5-minute telephone conversations involving 679 participants from the Mid-Western states (MN=156, WI=105, OH=70, IA=64, MI=41, IL=37). Participants in Phase II were recruited from the following college campuses: Iowa State University, Michigan State, University, University of Michigan, University of Minnesota, University of Wisconsin at Madison, Northwestern University, and Ohio State University. Switchboard 2 Phase 3 recruited in the American South with a balance of males and females.

During the first Switchboard Cellular collection (1999-2000), our goal was to collect 10 6-minute calls from 190 GSM cell-phone users balanced by gender. Our most successful means of recruitment was among the employees of a local GSM provider. However, we were still only able to recruit 293 participants, not enough to compensate for participant non-compliance. In a Switchboard study, participant non-compliance generally means that participants either do not make calls or do not receive calls. In the case of Switchboard Cellular, we found that the phones of many of many participants were unavailable during those times the participants themselves had agreed to receive calls. We discovered that this was a result of participants' habit of turning off their cell phones when not using them. To counter this problem and to reduce the frustration of those who subject who were initiating call, LDC recruiters contacted participants on multiple occasions via phone and mailings to remind participants to: 1) change their availability schedule to better reflect those times when they really could receive calls, 2) leave their phones on during the their "available" times, 3) be sure to initiate calls. In a further effort to encourage participation, we instituted a participant lottery for those who completed the study (10 calls). While this study proved to be one of our most efficient in terms of the number of subjects who finished the study, it was also our most labor-intensive study. To counter this participant non-compliance, we decided to over-recruit for the next phase of Switchboard Cellular conducted in the Fall 2000.

The goal in Switchboard Cellular 2 was to collect 10 calls each from 210 participants balanced by gender but with no restriction on cellular networks represented. We recruited a total of 591 participants and instituted a sliding pay scale that 1) covered subject costs for each call 2) offered a large bonus for completing the study (the 10th call) and 3) offered smaller bonuses for participation after the subject had completed the required number of calls. These measures provided strong motivation to the subjects and minimized the LDC's investment in under-performers. As a result, we were able to complete Switchboard Cellular Phase 2 in about one month.

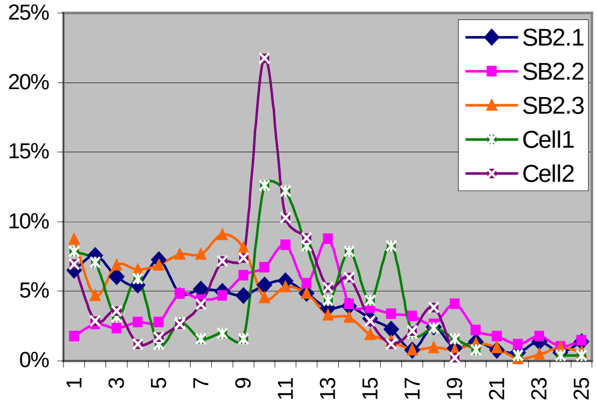


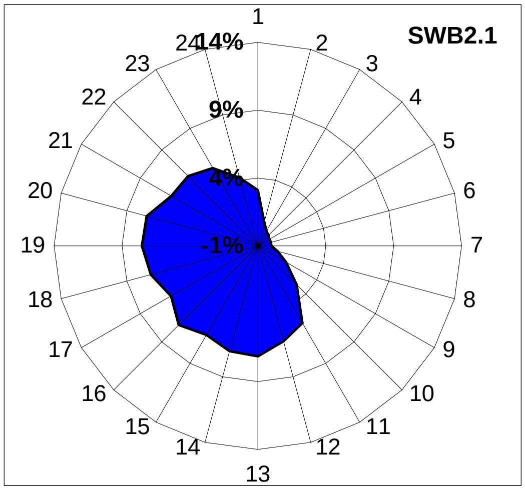
Figure 2 shows participant behavior in each of the Switchboard collections. Switchboard Cellular 2 has the tightest distribution of subjects around the goal of 10 calls. For Switchboard 2.1 through 2.3, the goal was to collect an average of ten calls per participant. Although these studies eventually met their goals, Figure 2 reveals a very diffuse distribution of participant performance. In the Cellular studies, the goal became having a minimum number of subjects who participated in at least 10 calls. The labor intensive approach adopted in Switchboard Cellular 1 produced a funny distribution of subject performance and was costly in terms of recruiter effort. The approach used in Switchboard Cellular 2 produced a distribution that is very tightly centered around a mode at ten calls and was in every other way, more efficient. Figure 2: Switchboard call summary. The vertical axis shows the number of participants who made the number of calls on the horizontal axis.

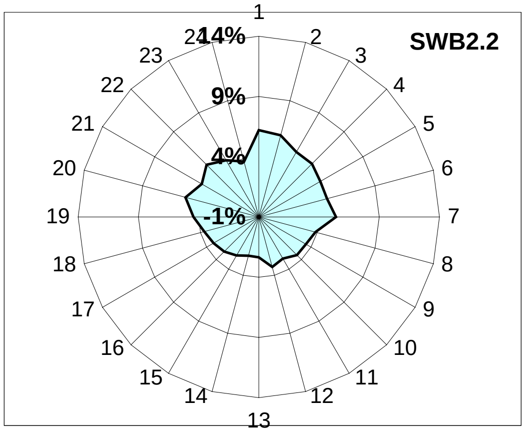
Figure 3 compares five Switchboard studies showing the ratio of completed subjects to total subjects and the normalized costs overall, per call and per subject who completed the study. The columns in Figure 3 refer to five Switchboard phases. "C1" and "C2" mean Switchboard Cellular Phase 1 and 2 respectively. The rows show the number of subjects, the subset of those who completed 10 calls, the ratio of those two numbers and the total number of calls collected. The second set of rows gives costs associated with participant fees for each of these studies. These costs are normalized for inflationary increases in the base compensation rate. The overall costs were first normalized for inflation and then scaled with respect to Switchboard 2 Phase 2, the largest and therefore most expensive collection. Cost per call was calculated by taking overall costs due to participant fees, normalizing them for inflation and then dividing by the number of successful calls. The cost per completed subject is the inflation-normalized sum of costs due to participant fees divided by the number of subject who completed 10 calls. These last two rows show that problems in the Switchboard Cellular Phase 1 collection caused an 8% increase in the cost per call relative to previous studies. However Cellular Phase 1 was relatively efficient with respect to completed subjects. Switchboard Cellular Phase 2 is the most cost-efficient study by both measures. The sliding scale compensation does seem to have had the effect of encouraging subject in Switchboard Cellular Phase 2 to make all of their 10 calls.

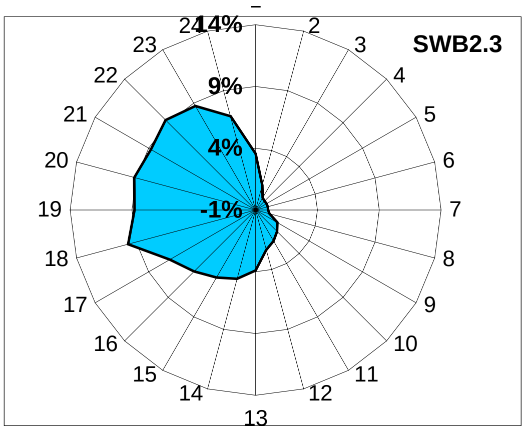
	2.1	2.2	2.3	C1	C2
Subjects	661	684	640	254	418
Completed Subjects	314	463	216	170	261
Ratio Subjects/Completed	0.48	0.68	0.34	0.67	0.62
Calls	1189	1322	1150	462	780
Normalized Participants Costs					
Overall	0.81	1.00	0.60	0.31	0.43
Cost/Call	1.00	1.00	1.00	1.08	0.97
Cost/Completed Subject	0.93	0.78	1.00	0.66	0.59

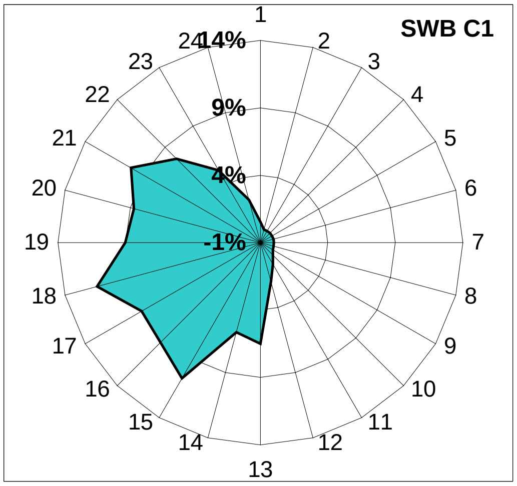
Figure 3: Summary of five Switchboard collections and their normalized costs.

The five Switchboard studies reported here all differ with respect to the hours of the day during which calls were completed. The policy established for Switchboard 2 Phase 1 and retained in Phase 2 allowed calls between the hours of 12:00 noon and 2:00AM. In Phase 3, the schedule was abbreviation to noon to midnight. In each of these studies, however, this was merely a policy statement, the robot operator was available around the clock and subject could give hours of availability outside the suggested times. In the Cellular phases, we converted our recommendation into a hard limit. Generally, the robot operator would not accept calls and recruiters would not enter availability times after midnight and before noon. We believed these time restrictions were necessary in the smaller Cellular studies to ensure that there was a critical mass of participants making and receiving calls during a smaller block of time. Figures 4 through 8 show call activity as a function of hour of the day in each of five Switchboard studies. The 24 hours of the day run clockwise around the circumference of the graph. For each hour of the day, the percentage of successful calls made during that hour is plotted as distance from the center of the circle. The points are then connected to form an area graph. These graphs are described below.









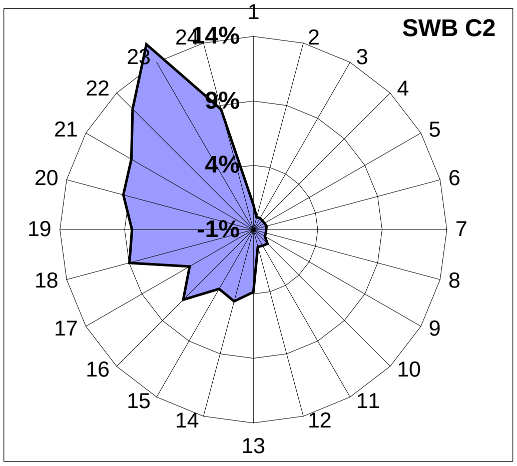


Figure 4-8: Call activity as a function of hour of day in five Switchboard collections

In Switchboard 2 Phase 1 call activity is concentrated in the time between 11:00AM and 1:00AM. In Phase 2, there is more call activity outside the suggested hours than there is within. It is important to note here that we are reporting local time, that is Eastern Standard Time. Many participants in Phase 2 were calling from and to areas that were one or two time zones west of the robot operator. Even so, Phase 2 stands out for the number of late night activity. Phase 3, focussed on the American South does return to the calling pattern of Phase 1. Indeed calls are more tightly clustered in the hours between 2:00PM and 11:00PM. In the Cellular phases, there is essentially no activity between midnight and noon. Note however, that the distribution of calls by hour of the day is different in the two studies. Many of the participants in Cellular Phase 1 were employee of a local call center for a cellular phone service provider. One can observe peaks of activity at 1:00PM, 3:00PM and 6:00PM and 9:00PM that presumably correspond to lunch, break-time, end-of-day and the hour after dinner. In Cellular Phase 2 we see similar peaks and one additional peak at 11:00 PM just before the robot operator shut down for the day.

By analyzing call activity in the previous studies, we were able to adjust the rules of Cellular Phase 2 participation to increase the chance that a caller would quickly connect with a callee.

Comments from subjects collected after the studies ended, reflect the difference between Cellular Phases 1 and 2. In the former, many subjects commented that it was difficult to find someone to talk to. Such problems all but disappeared in Cellular Phase 2.

7. Outcomes

Speech data is crucial to speech technology. This paper compared methodologies for collecting speech data across more than two-dozen studies showing LDC procedures for several collection types and the relative efficiency of various approaches. This section will enumerate some of the resources that these efforts have produced including those that are generally available and those that will be. All for the data described herein is or will shortly be available for research and technology development. The parenthetical notes after each corpus name are the LDC Catalog number and the ISBN number

The CallFriend corpora created for language identification are available in American English-Non-Southern Dialect (LDC96S46 isbn:1-58563-061-6), American English-Southern Dialect (LDC96S47 isbn:1-58563-062-4), Canadian French (LDC96S48 isbn:1-58563-063-2), Egyptian Arabic (LDC96S49 isbn:1-58563-064-0), Farsi (LDC96S50 isbn:1-58563-065-9), German (LDC96S51 isbn:1-58563-066-7), Hindi (LDC96S52 isbn:1-58563-067-5), Japanese (LDC96S53 isbn:1-58563-068-3), Korean (LDC96S54 isbn:1-58563-069-1), Mandarin Chinese-Mainland Dialect (LDC96S55 isbn:1-58563-070-5), Mandarin Chinese-Taiwan Dialect (LDC96S56 isbn:1-58563-071-3), Spanish-Caribbean Dialect (LDC96S57 isbn:1-58563-072-1), Spanish-Non-Caribbean Dialect (LDC96S58 isbn:1-58563-073-X), Tamil (LDC96S59 isbn:1-58563-074-8), Vietnamese (LDC96S60 isbn:1-58563-075-6)

The CallHome corpora were created for large vocabulary continuous speech recognition. For each CallHome language, LDC has released audio, transcripts and a lexicon. Due to space constraints, the identifying information for only the audio follows. The LDC Catalog (www.ldc.upenn.edu/Catalog) gives links from the audio to the transcripts and lexicons. CallHome corpora are available in the following: American English (LDC97S42 isbn:1-58563-111-6), Egyptian Arabic (LDC97S45 isbn:1-58563-114-0), German (LDC97S43 isbn:1-58563-117-5), Japanese (LDC96S37 isbn:1-58563-077-2), Mandarin Chinese (LDC96S34 isbn:1-58563-080-2) and Spanish (LDC96S35 isbn:1-58563-083-7).

The Hub-5 training and evaluation corpora for large vocabulary continuous speech recognition are available in: Mandarin (LDC98S69 isbn:1-58563-131-0) and Spanish (LDC98S70 isbn:1-58563-133-7)

Switchboard 1 (LDC97S62 isbn:1-58563-121-3), three phases of Switchboard-2 Phase 1 (LDC98S75 isbn:1-58563-138-8), Phase II (LDC99S79 isbn:1-58563-144-2), Phase III (LDC2002S06 isbn:1-58563-222-8) and one Switchboard Cellular Phase 1 (LDC2001S13 isbn:1-58563-213-9) have been published. LDC has also released a subset of transcribed Switchboard Cellular calls both audio (LDC2001S15 isbn:1-58563-215-5) and transcripts (LDC2001T14 isbn:1-58563-214-7).

NIST's Speaker Recognition Benchmarks for 1996 (LDC96S61 isbn:1-58563-059-4), 1997 (LDC99S80 isbn:1-58563-142-6), 1998 (LDC98S76 isbn:1-58563-129-9), 1999 (LDC99S81 isbn:1-58563-152-3) and 2000 (LDC2001S97 isbn:1-58563-192-2) have all been released.

To date about 30 hours of meetings involving more than 90 unique speakers have been collected under the GroupTalk and GroupMeet protocols. Some of this material will be included in NIST's Rich Text 2002 Metadata Annotation Experiment.

8. References

[1] Doddington, George, 1998, The Topic Detection and Tracking Phase 2 (TDT-2) Evaluation Plan: Overview & Perspective, Proceedings of the Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, February 1998.

[2] Garofalo, et. Al., 2000, The TREC Spoken Document Retrieval Track : A Success Story, April 2000.

[3] NIST - National Institute for Standards and Technology, 2002, Rich Text 2002 Metadata Annotation Experiment http://www.nist.gov/speech/tests/rt/rt2002/experiment.htm

[4] NIST - National Institute for Standards and Technology, 1999, 1999 NIST Broadcast News Evaluation, http://www.nist.gov/speech/tests/bnr/bnews_99/bnews_99.htm

[5] NIST - National Institute for Standards and Technology, 2000, ACE - Automatic Content Extraction, http://www.nist.gov/speech/tests/ace/

[6] NIST - National Institute for Standards and Technology, 2000, The 2000 NIST Hub-5 Evaluation, http://www.nist.gov/speech/tests/ctr/h5_2000/index.htm

[7] Wayne, Charles, 1998, Topic Detection & Tracking: A Case Study in Corpus Creation and Evaluation Methodologies, Proceedings of LREC 1998: The First International Conference on Language Resources and Evaluation, Granada, Granada, Spain, May 1998.

[8] Wayne, Charles, 2000, Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, Proceedings of LREC 2000: The Second International Conference on Language Resources and Evaluation, Athens, Greece, May 2000.