SWITCHBOARD: Telephone Speech Corpus for Research and Development

John J. Godfrey Edward C. Holliman Jane McDaniel

Texas Instruments, Inc., Dallas, TX 75265

Abstract

SWITCHBOARD is a large multispeaker corpus of conversational speech and text which should be of interest to researchers in speaker authentication and large vocabulary speech recognition. About 2500 conversations by 500 speakers from around the U.S. were collected automatically over T1 lines at Texas Instruments. Designed for training and testing of a variety of speech processing algorithms, especially in speaker verification, it has over an hour of speech from each of 50 speakers, and several minutes each from hundreds of others. A time-aligned word for word transcription accompanies each recording.

1 INTRODUCTION

The appetites of today's statistical speech processing techniques for training material are well described by the aphorism: "There's no data like more data." Large structured collections of speech and text are essential to progress in speech and speaker recognition research. The series of DARPA-sponsored corpora most often referred to as TIMIT [1][2], Resource Management [3], and ATIS [4], for example, have become important landmarks in the field of continuous speech recognition.

In this paper we describe a DARPA-sponsored corpus of spontaneous conversational speech which addresses the growing need for large multispeaker databases of telephone bandwidth speech. The SWITCHBOARD corpus, collected at Texas Instruments, is particularly valuable as a rich and varied source of speaker authentication data sets.

2 SWITCHBOARD CORPUS

SWITCHBOARD, which is about 65% complete at the time of this writing, will include 2500 conversations of three to ten minutes' duration, carried on by about 500 paid volunteers of both sexes from every major dialect of American English. In round numbers, this amounts to over 250 hours of speech and nearly 3 million words of text.

Apart from sheer volume, however, SWITCHBOARD has a number of unique features designed to support research in both speaker authentication and speech recognition for telephone-based applications. Among these features are:

- automated collection
- all-digital 4-wire format
- detailed transcription and time alignment of all conversations
- design for multiple testing runs and a variety of technical approaches
- an underlying relational database system

2.1 Automated Collection

The conversations in SWITCHBOARD were collected under computer control, without human intervention. From a human factors perspective, automation guards against the intrusion of experimenter bias, and provides a degree of uniformity in the collection environment. There is a potential disadvantage, in that once an automated system is in place, it cannot readily be

0-7803-0532-9/92 $3.00 © 1992 IEEE

changed, even when experience shows it should be. The SWITCHBOARD collection protocol was therefore developed over months of extensive pilot testing. Experiments determined the kind and amount of man-machine interaction, the order of presentation, the types of prompts, and other factors needed to insure smooth participation and to elicit natural and spontaneous speech by the participants.

The hardware platform was a commercial system known as a "Robotoperator," consisting of an IBM Model 80 computer, 700MB disk drive, T1 interface, and a programmable switching system for connecting among the channels of the T1 span.

An application software package intended for commercial use was modified to control the SWITCHBOARD protocol. Upon receiving an incoming call on one of the T1 channels, it would play appropriate messages, and collect touchtones indicating the caller's identification and telephone number. A resident database management system attempted to find other participants who had not spoken with this caller or heard this prompt before, and who could be reached at this time. The system then called out on another of the T1 channels until one of the prospective partners responded and indicated he was ready to participate.

The instructions include a prompt suggesting a topic of conversation, and a reminder that the talkers are free to hang up at any time, or to take as much time as they wish to introduce themselves and "warm up" before giving a signal to begin recording.

Transcribers were later asked to rate the naturalness of conversations they listened to, using a five point scale from "very natural" (1) to "forced or artificial-sounding" (5). The average rating of 1.48 suggests that the collection protocol was successful in this respect.

2.2 All-digital 4-wire Format

The use of a dedicated T1 line, a customized computer interface, and automatic switching software made it possible to collect the digital version of the speech signals directly from the telephone network, and even to record the two sides of each conversation separately but synchronously. The T1 line bypassed local Central Office switches on the way to the MCI long distance network, eliminating all possible sources of analog conversion at TI's end of the calls. The goal was real telephone speech, but of the best available quality, with no degradation due to the collection system.

Before the two callers are interconnected, they are engaged in two separate telephone calls with the computer, listening to recorded messages and sending touchtones back. Once this phase is complete, the T1 controller connects caller A's "Transmit" side to caller B's "Receive" side, and vice versa, and at the same time passes the two "Transmit" signals to the computer interface, where they are written to disk as separate incoming messages. This pair of 8 kHz, mu-law encoded signal files is later integrated into one file in NIST's standard SPHERE format, with alternating bytes from the two sides of the conversation interleaved. The released version will contain software for accessing either side or the sum of the two sides.

The isolation of the two talkers is limited by the long distance telephone network's echo cancelling performance, but is generally better than 20dB. Many calls have no audible "echo" on either side. This should facilitate training separately on an individual's voice, and then testing algorithm performance on either one or both speakers from a given conversation.

2.3 Time-aligned Transcription

Each conversation is fully transcribed, with special conventions to show speakers' turns, simultaneous talking, interrupted sentences, partial words, and other phenomena common in spontaneous conversational speech. After producing the text, the transcribers were asked to rate each conversation on a number of properties, such as the amount of background noise or static, difficulty in understanding the talkers, and degree to which the conversants stayed on one subject. A standard set of terms to describe background noises, and a format for inserting comments in the text, were

also provided.

The SWITCHBOARD conversations are also time aligned at the word level. The time alignment is accomplished using supervised phone-based recognition, as described in a companion paper by Wheatley [5]. This process produces phone by phone time markings, which are then reduced to a word by word format for publication with the transcripts. The original phonetic base forms will be available in a dictionary with the SWITCHBOARD corpus. In general, this process specifies the endpoints of words to within a few tens of milliseconds, except in a few problem areas such as simultaneous speech. See the Wheatley paper for details.

The corpus is therefore capable of supporting not only traditional text-independent approaches to speaker verification, but also those which make use of some form of knowledge of the text, such as speaker dependent variations in specific phonetic models. Furthermore, in spite of the bandwidth limitations, the signal quality of many of the conversations is high enough that SWITCHBOARD data may serve as a platform for a variety of important experiments in large vocabulary speech recognition. The texts alone should have significant value in developing statistical language models for spontaneous conversational speech.

2.4 Corpus Design

The design of the SWITCHBOARD corpus emphasizes the importance of both depth and breadth of coverage, especially for speaker verification research. Fifty "target" speakers participated at least 25 times, which adds up to more than an hour of speech gathered over a period of weeks. The target callers were no different from the rest of the population, except that they had to make and receive calls with multiple telephone instruments. This was intended to prevent speakers from being identified by channel effects due to handset characteristics.

The design assumes that 25 target speakers should suffice to get statistically reliable estimates of the performance of a speaker verification algorithm under development; an equal number is available to be set aside for final evaluation of the system. The amount of training data needed per speaker is harder to predict, but most systems would probably use no more than half of the 60 to 90 minutes to be found in SWITCHBOARD. The remainder should then be enough to support numerous re-tests on previously unseen material from each talker during algorithm development. The configuration illustrated in Figure 1 shows 60% of the conversations being used for training and 40% for repeated tests.

	Sessions					
	Training	Test Sets				
Spkr	1 2 ... 15	16	17	...	25	Size
1	x x ... x	x	x	...	x	
2	x x ... x	x	x	...	x	625
⋮						Develop-ment
25	x x ... x	x	x	...	x	tokens
26	x x ... x	x	x	...	x	
27	x x ... x	x	x	...	x	625
⋮						Evalua-tion
50	x x ... x	x	x	...	x	tokens
51	x x ... x					
52	x x ...					
⋮		Imposter Set(s)				
i	x x					
$i+1$	x					
⋮		3750 tokens				
499	x					
500	x					

Figure 1: A possible configuration of SWITCHBOARD for speaker authentication. Each token (x) is one talker's side of one conversation.

The other 450 speakers who participate in from one to twenty calls constitute a pool of "imposters." While this may fall short of the ideal of one call per caller for thousands of callers, whereby every imposter call is uncorrelated with every other, it is certainly large enough to support a variety of open-set experimental designs. One could imagine dividing the callers into devel-

opment and evaluation cohorts, or mixing callers and dividing their conversations, or using all of their conversations in all experiments.

The amount of material and the number of speakers also makes it possible to run both speaker independent and speaker dependent experiments in large vocabulary speech recognition. Although the speech is conversational and spontaneous, seventy different prompts were used to stimulate the conversations. This will undoubtedly produce natural groupings in terms of lexical coverage, and likely other linguistic factors as well.

2.5 Underlying Database

Demographic information about the speakers is entered in an Oracle data base when they register to participate. This includes their age, sex, level of education, and the geographically-defined dialect area where they grew up. The callers' identification numbers, the date, time, and length of the call, as well as the area codes and telephone numbers of the participants and other pertinent information about each call are all automatically entered into Oracle tables.

The information in these tables, except for protected personal data, will accompany the speech and text files when the SWITCHBOARD corpus becomes publicly available. Distribution will be through the National Institutes of Standards and Technology.

This work was sponsored by DARPA/SPAWAR Contract No. N00039-90-0168.

References

[1] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proceedings of the DARPA Speech Recognition Workshop*, 1986.

[2] L. Lamel, R. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proceedings of the DARPA Speech Recognition Workshop*, 1987.

[3] Price, P.J., W.M. Fisher, J. Bernstein, D.S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition", *Proceedings of ICASSP*, 1988.

[4] Charles T. Hemphill, John J. Godfrey, and George R. Doddington, "The ATIS Spoken Language Systems Pilot Corpus," *Proceedings of the DARPA Speech and Natural Language Workshop*, Pittsburgh, PA, June, 1990.

[5] B. Wheatley, G. Doddington, C. Hemphill, J. Godfrey, E. Holliman, J. McDaniel, and D. Fisher, "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech," *Proceedings of ICASSP*, 1992.