THE SUMMERS OF OUR DISCONTENT

Jordan R. Cohen

Center for Communication Research

Thanet Road

Princeton, NJ. 08540

ABSTRACT

A series of summer workshops has been held over the past four years. The topics have focused on speech recognition applied to the Switchboard corpus.

This note offers a short historical perspective on the data, the topic, and the first two workshops. The title is not an expression of unhappiness, but rather a reminder that sensible research focuses on unsolved problems.

1. IN THE BEGINNING

In the decade of the 80's and the early '90s, ARPA was the major funding source for research in speech recognition in the United States. Problems for recognition research progressed from isolated words to continuous speech, and two corpora engendered most of the research effort: ATIS and Wall Street Journal.

The ATIS corpus is a collection of conversations between a travel planner and an artificial computerized airline information system. In the early stages, the systems were simulated with a "man behind the curtain", while in the later stages data was collected using real systems with real results and honest errors.

The second corpus, the Wall Street Journal, was a large collection of data read from articles in the newspaper. The speech was collected in a controlled acoustic environment, and performance flirted with word error rates in the single digits. The problem seemed in hand.

Early in the 90's, the Switchboard corpus was collected by Texas Instruments. Its corpus of telephone conversations were meant to be used to advance the field of "topic spotting". The conversations were unrehearsed discussions between strangers using standard telephones, and recorded from long distance phone lines. Early work by Dragon Systems on this corpus using their speech recognition system yielded errors in the high 70 percent range. A new era in speech recognition research had begun.

2. THE IDEA

I was concerned that the research community might continue with the Wall Street Journal, ignoring the telling results from Switchboard. To that end, I recommended to ARPA that we hold a summer workshop (a working meeting, like those held each year at IDA) with a focus on the Switchboard corpus. The ARPA program manager was interested, as were some Department of Defense agencies. I was offered half of the anticipated $1,000,000 cost from each. Ultimately, ARPA was not able to fulfill its offer in cash, but it did recommend that it's contractors attend the workshop, and allowed the workshop expenses as part of the ARPA research projects.

2.1. Frontiers in Speech Processing I: Robust Speech Recognition '93

The Center for Computers in Aid to Industrial Productivity at Rutgers offered to host the first workshops. Jim Flanagan served as host to the first and second workshops, and his capable staff made dealing with computers, networking, and living situations manageable.

The 1993 workshop had several positive aspects. High capacity workstations had become available at a reasonable cost (in our case, Sparc 10's). The Switchboard data had been released, and BBN was able to provide us corrected transcripts, a pronouncing dictionary, and other data for a 4 hour subset of the corpus known as the "credit card corpus". Finally, Cambridge University and Entropic Systems had a capable hidden Markov model toolkit, HTK, which we were able to provide as a standard working model.

Entropic Systems produced a "baseline" recognition system before the workshop began, although we were debugging through the first few weeks of our meeting. The original system produced a 79% word error rate before the summer, and with final debugging this was lowered to about 75%.

Several ideas led to lowered error rates during the summer. Smoothing the language model yielded a few percent, more complicated acoustic models yielded a few percent, and several ideas (analysis-by-synthesis, time/frequency kernels, and adaptive models) looked promising but were not able to be tested in the time and computing that we had available. Splitting the data into Male and Female corpora yielded the best results (67% error), and were produced by both the BBN team and the SRI researchers, each using the internet to do experiments with their home systems.

- Several experiments suggested that we had not hit a structural limit:
- Recognition of the training material yielded 24% error.

- Hillclimbing using the TEST material to adjust the means of the Gaussians in the acoustic model yielded 30% error.

This first workshop also yielded several side results:

- A software package was produced to add "realistic" noise to speech, as suggested by Jont Allen's continuing discussion of Fletcher's work.
- A very successful speaker identification system was produced using Gaussian models and high-order LPC observations.
- RASTA signal processing software was written, documented, and made available to the community.
- A new software package for computing time/frequency displays was written and made available.

The workshop was judged a success, and a CD of the notes of the participants, the data, and any results from this workshop were made available through NIST and the LDC [1]. (This CD also contains a digital copy of the RASTA song - don't miss it!).

2.2 Frontiers in Speech Processing II: Switchboard, the Second Year

The second summers' workshop profited from our experiences during the first summer. Living conditions improved. We organized "interest groups" for focus, and we bought not only Sparc 5 workstations, but, with ARPA's assistance, we bought the "computing cluster" of 20 Sparc 20's for use as compute service for the workshop. The membership grew from 25 to 32, and the performance improved, although not as quickly. Herb Gish served as co-chair.

The focus was broadened to include the "10-topic" subcorpus of Switchboard. During the intervening year, the community had demonstrated performance of almost 50% correct recognition using 80 hours of acoustic training: the performance was noticeably worse if only 25 hours' training was used, and at 4 hours performance mimicked the 1993 workshop error rates.

The six focus teams in this workshop were:

- Speaker Variability
- Language Modeling
- Consistency of Transcription and Data
- Evaluation and Diagnostics
- Mathematical Modeling
- Channel Effects

Early in the summer, the Transcription and Evaluation groups merged, since they had common membership and interests.

This second summer, unlike the first, was filled with negative results using clustering algorithms, new front end applications and noise reduction algorithms. Each of the algorithms tried had either a current guru or it's original author in attendance at the workshop, so we have a reasonable confidence that the experiments would have yielded results had there been an effect. There was essentially none.

The transcription group worked on "shorties", or those words whose representation in the final answer was only a few frames long. The preponderance of short segments was cleared up by rewriting the dictionary for commonly occurring words, but the accuracy remained unchanged. Work was done on extensions of HMM models, and on rate effects.

There were only two positive effects to be reported. The first was in language modeling. Modeling common word pairs as dictionary entries, a "bigram" grammar was produced which actually spanned four words at a time for commonly occuring strings. This grammar lowered the error rate by 2%.

Speaker adaptation in terms of vocal tract length modeling (defined as stretching or squeezing the spectrum of the speech) was demonstrated to have a significant effect on error rates. Evaluations showed 3-4% improvements, absolute. You may see the echoes of this work in many of the current day systems.

As before, a CD was published with the data, notes, and results of the workshop. It, too, is available from the LDC [2].

3. THE DATA

It might be useful to listen to a few sentences from Switchboard. It is difficult to imagine a more "natural" discourse style, and the challenges of recognition on these data are many. My kudos to those members of the early workshops who brought us along to the current situation, and to the standard bearers who did work throughout the years on this corpus, and shared their results with us.

Here are a few examples of the real speech from Switchboard:

4. SUMMARY

In summary, I'd like to say how much I enjoyed having a part in making these workshops happen. It was exciting to see standard solutions fail on this challenging corpus, while novel ones began to take shape. Our international participants spread the excitement to other countries, and the Switchboard evaluations, now known as Hub 5, have multinational participation. There were many small projects which I didn't have time to discuss here, and I encourage you to read the CD records. Take from them what you find interesting.

5. REFERENCES

1. Frontiers in Speech Processing, The 1993 CD-ROM, LDC LDC96LS28.
2. Frontiers in Speech Processing, The 1994 CD-ROM, LDC LDC96L40