INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

Improved Monosyllabic Word Modeling on SWITCHBOARD

February 15, 1999

WE WON THE BID TO CREATE A DIGITAL ARCHIVE OF THE WORLD'S GREATEST ART.
www.dilbert.com
THIS WILL GIVE US A CHANCE TO FIX ANY ERRORS MADE BY THE ARTISTS.
ERRORS?
FOR EXAMPLE, THERE WAS A GUY WHO USED TOO MUCH BLUE FOR A WHOLE PERIOD.

submitted by:

V. Mantha, J. Hamaker, N. Deshmukh, A. Ganapathiraju, and J. Picone
Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
Box 9571
413 Simrall, Hardy Road
Mississippi State, Mississippi 39762
Tel: 601-325-3149
Fax: 601-325-3149
email: {mantha, hamaker, picone}@isip.msstate.edu

EXECUTIVE SUMMARY

SWITCHBOARD (SWB) Corpus consists of 2438 conversations digitally recorded over long distance telephone lines. The SWB Corpus totals over 240 conversation hours (elapsed time) of data. The average conversation duration is six minutes. The transcriptions contain more than 3 million words of text. The SWB Corpus includes more than 500 adult-aged speakers and covers most major American English dialects. Such impressive statistics make SWB the premier database for telephone bandwidth large vocabulary conversational speech recognition (LVCSR) research. The goal of this project is to resegment the speech data and correct the transcriptions in an effort to significantly advance LVCSR technology.

This project has reached its midpoint. Thus far, we have released 367 conversations with corrected segmentations and transcriptions, and another 551 conversations with complete segmentations. These conversations comprise 50% of the conversations used in the WS'97 partition and 40% of the entire SWB corpus. We have also demonstrated that one could obtain a 2% decrease in WER by simply reestimating LVCSR models on the corrected segmentations. This work was presented at the recent Hub-5 Conversational Speech Recognition (LVCSR) Workshop and was met with much support as well as many suggestions for improvement from the speech research community. Keeping these suggestions in mind, we have made numerous revisions to our segmentation and transcription procedures. These include

- minimizing the amount of non-speech data included in an utterance definition. We are also attempting to maintain as much phrase structure as possible while still allowing for silence paddings;
- a new workflow process in which we do segmentations first, and then provide transcriptions in a second pass through the data. We plan on completing the segmentations by March 31, 1999. To this end all our validators are presently working on segmentations.

We have also implemented an incremental and multiple-pass quality control procedure which provides almost immediate feedback to the validators. For a database of the magnitude of SWB, it is imperative that the quality checks be done almost immediately after the data has been segmented. This approach has led to a decrease in the error rate as well as an increase in productivity. We had earlier reported a cross-validation of less than 1% WER for the transcriptions of a relatively clean utterance. These figures show a substantial improvement over the current LDC transcriptions which have an 8% WER measured under the same conditions.

By March 31, we will release the final segmentations for all 2438 Switchboard conversations. This will be a major milestone in the project, and be a sign the end is in sight. We then will turn our attention to transcription correction, and project a completion date of August 31, 1999. Such a timeframe will allow us to complete the manual and automatic word alignment generation and review before the projected deadline of December 1999. All information relevant to this project is also available at *http://www.isip.msstate.edu/projects/switchboard/*, including the most current set of segmentations.

Our Switchboard mailing list has recently grown to 23 users, an indication of the renewed interest in this project and its data. Discussions are underway to support several projects interested in using the data for tasks ranging from phonetic recognition to prosodic labeling. Reconciliation of the data with other SWB resources is becoming increasingly more important.

1. ABSTRACT

In our last report [1] we described an improved workflow process based on a new set of segmentation and transcription guidelines. These new guidelines were a result of numerous suggestions received from colleagues during the initial months of this project. We also implemented an incremental and multiple pass quality control procedure. Cross-validation tests showed that these new procedures decreased our WER from 3% to less than 1% (the original LDC transcriptions had a WER of 8% measured under the same conditions).

In this quarter we have not made any changes to the segmentation and transcription conventions. We did, however, make a few adjustments to our work procedures to counter abrupt changes in staffing. Key among these is our decision to only segment data (no transcriptions are changed) until the entire corpus has been completely segmented. This new procedure gives our new validators the chance to be productive almost immediately. We plan to complete segmentation of the entire database by March 31 and then shift our focus to transcriptions. Though we are slightly behind schedule, the projected completion date of this work remains at December 1999.

2. INTRODUCTION

The SWITCHBOARD Corpus [2, 3] has become critical to the success of state-of-the-art LVCSR systems. Using this data, however, has not been without its share of drawbacks. SWB was a great example of the trials and tribulations of database work in that the quality of the data suffered from a lack of understanding of the problem. Word-level transcription of SWB is difficult, and conventions associated with such transcriptions are highly controversial and often application dependent. By 1998, the quality of the SWB transcriptions for LVCSR was recognized to be less than ideal, and many years of small projects attempting to correct the transcriptions had taken their toll. Numerous versions of the SWB Corpus were floating around; few of these improved transcriptions were folded back into the LDC release; and many sites had spent a lot of research time cleaning up a portion of the data in isolation. In February of 1998, ISIP started a project to cleanup the SWB Corpus, and to organize and integrate all existing resources related to the data into this final release.

In the first six months of this project, we made considerable progress in transcribing and resegmenting the corpus. We released 1000 of the 2438 SWB conversations with revised segmentations and transcriptions. As noted in our last report, though, we revised those conversations to conform to some very important conventions [4] (marking of boundaries near noise for example). We also amassed a large collection of tools and resources for use with the SWB project. Most notable of these are the development of our public-domain segmentation tool [5], the SWB frequently asked questions (FAQ) web-site [6], the SWB educational resources web-site [7], and a comprehensive collection of statistics [8] related to SWB. We continue to maintain a mailing list (*swb@isip.msstate.edu*) which is our point of contact to the research community for resolving subtle transcription issues and communicating progress on our efforts.

Our plan at that point had been to continue working on segmentation and transcription simultaneously, finishing the SWB Corpus by the beginning of August 1999. However, in the last three months of this work we have been forced to alter our previous plans. This was done to account for changes in staffing and to make up for the time spent in training of the new validators. These changes and the adjustments made in our schedule are described in the next sections.

3. CHANGES IN WORKFLOW PROCESS

To insure that the released data is in accordance with the transcription and segmentation conventions, it is important that the validators' work be reviewed almost immediately after it is completed. This provides the validator with immediate feedback and allows the project manager to prevent any long-term problems in the data. In the previous quarterly report we described a process wherein some validators were working on segmentations and others were focusing only on transcriptions. However, staffing changes have forced us to reconsider this strategy.

To get our new validators up to speed quickly, we have decided to allow them to focus on the relatively simple task of segmentation. It is our plan to first complete segmentation of the entire corpus and then to again churn out revised transcriptions. A majority of the effort spent on this work is in transcription and word alignment review so the amount of time required to complete segmentation of the corpus is minimal. Yet it is sufficient to allow our newer validators to gain experience before beginning transcriptions.

Limiting our validators to only segmentations provides two other advantages. First, the research community will have the opportunity to begin working with the new segmentations across the entire corpus. It is our opinion that this will better serve the community than waiting for incremental releases of corrected segmentations and revised transcriptions. The other advantage is that this lets our new validators be productive almost immediately. The average training time is less than one week for segmentation; training time for transcriptions is nearly triple that.

4. RELEASED DATA

Since our last progress report, we have released 162 more conversations with revised transcriptions and 552 having only revised segmentations. We have also released the complete Switchboard database (with a mixture of LDC/ISIP transcriptions) to satisfy requests from members of the research community. In total, we have released 362 conversations with completed segmentations and transcriptions and another 556 conversations with complete segmentations. This comprises 151 hours of conversation data. Another 250 conversations will be released shortly. The released data covers the WS'97 devtest set, WS'97 eval set and part of the WS'97 training set. Table 1 gives some pertinent details regarding the released data.

We will continue to make incremental releases of newly segmented data available to the community via our project web page. A revised timeline for these releases is shown in Figure 1. We will make a release of the entire SWB Corpus with revised segmentations on March 31.

Conversations	918
# of non_silence utterances	71948
# of silence-only utterances	40452
hours of data	151.42
hours of speech data	93.09
Mean utterance duration	4.66

Table 1. An overview of the released data.

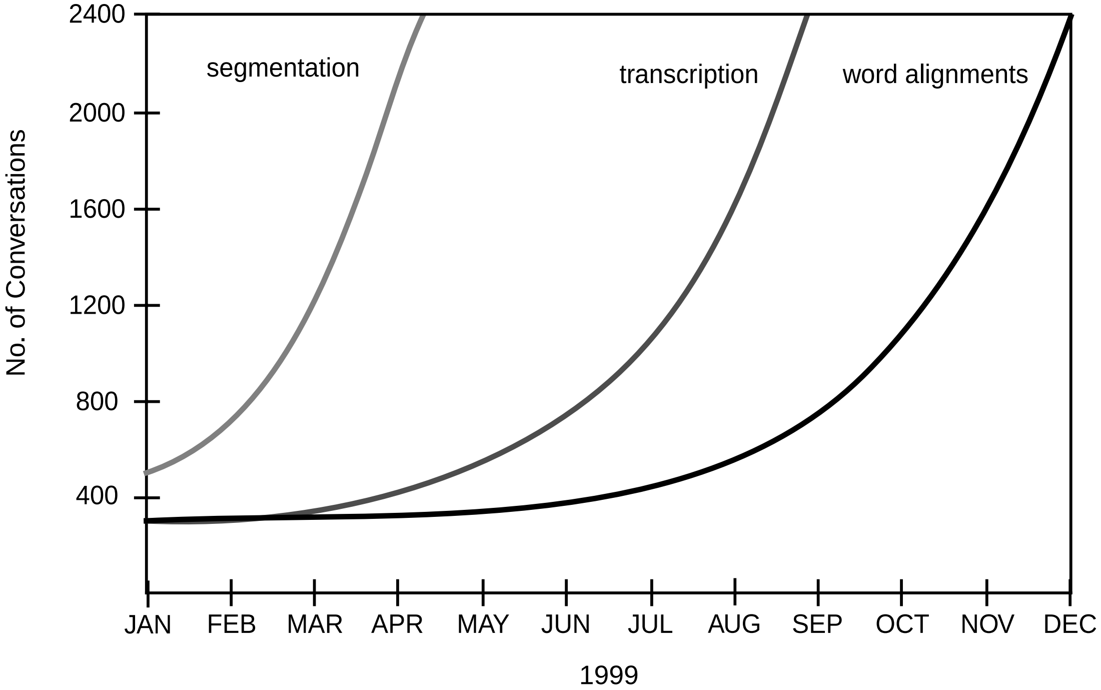

Figure 1. Timeline for the remainder of the SWB resegmentation project. Our anticipated completion date remains at December 1999.

Following this, we will begin to make incremental releases of the revised transcriptions. We expect the delay incurred in switching the new validators back to transcriptions to be minimal since each will have at least two months of experience at that point. A major release of the corpus with revised transcriptions and segmentations is set for early June. This is followed in December 1999 with a final release of the entire corpus including manual word alignments.

5. ISSUES

As this work continues it is clear that each validator reaches a point where they simply burnout. In this quarter we lost our two most experienced validators for this reason. Each had been with us for over six months and neither returned to work after the Christmas break. Of course, this practically stalled the project until we were able to hire and train a number of new validators. We fell approximately one month behind schedule during this period of training. The delay was limited since we were able to get the new validators into production mode quickly by allowing them to only work on segmentations. Our most successful validators associated with this project have survived about nine months. We have good reason to believe that one year is the upper limit of what we can expect from a good hourly student worker.

Though these setbacks were devastating at the time, it appears that the work is getting back on track. Our current set of validators are actually producing more data per week than the former validators. Our initial examinations of the new data also shows that the quality of the segmentations has not decreased with the new workers. If this trend continues, by late March we will again be on track for a December 1999 completion date.

6. ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our sponsor, the Department of Defense for their unwavering support of this project. We also appreciate the assistance of Dr. David Graff of the LDC for his help in various aspects regarding the SWB data.

7. REFERENCES

[1] J. Hamaker, N. Deshmukh, A. Ganapathiraju, and J. Picone, "Improved Monosyllabic Word Modeling," *Department of Defense*, August 15, 1998.

[2] J. Godfrey, E. Holliman and J. McDaniel, "Telephone Speech Corpus for Research and Development," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 517-520, San Francisco, California, USA, March 1992.

[3] B. Wheatley, G. Doddington, C. Hemphill, J. Godfrey, E.C. Holliman, J. McDaniel, and D. Fisher, "SWITCHBOARD: A User's Manual," *http://www.cis.upenn.edu/~ldc/readme_files/switchbrd.readme.html*, Linguistic Data Consortium, University of Pennsylvania, December 1995.

[4] J. Hamaker, Y. Zeng, and J. Picone, "Rules and Guidelines for Transcription and Segmentation of the SWITCHBOARD Large Vocabulary Conversational Speech Recognition Corpus," *http://www.isip.msstate.edu/resources/technology/projects/current/switchboard/doc/transcription_guidelines*, Institute for Signal and Information Processing, Mississippi State University, July 1998.

[5] N. Deshmukh, A. Ganapathiraju, R. Duncan, and J. Picone, "An Efficient Tool For Resegmentation and Transcription of Two-Channel Conversational Speech," *http://www.isip.msstate.edu/resources/technology/software/1998/swb_segmenter,* Institute for Signal and Information Processing, Mississippi State University, August 1998.

[6] J. Hamaker and J. Picone, "The SWITCHBOARD Frequently Asked Questions (FAQ)," *http://www.isip.msstate.edu/resources/technology/projects/current/switchboard/faq*, Institute for Signal and Information Processing, Mississippi State University, August 1998.

[7] J. Hamaker, A. Ganapathiraju, and J. Picone, "SWITCHBOARD Educational Resources," *http://www.isip.msstate.edu/resources/technology/projects/current/switchboard/doc/education*, Institute for Signal and Information Processing, Mississippi State University, August 1998.

[8] J. Hamaker and J. Picone, "A Statistical Guide to SWITCHBOARD: Topic Statistics," *http://www.isip.msstate.edu/resources/technology/projects/current/switchboard/doc/statistics*, Institute for Signal and Information Processing, Mississippi State University, August 1998.