Bi-Annual Status Report For

# Improved Monosyllabic Word Modeling on SWITCHBOARD


submitted by:

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: {hamaker, picone}@isip.msstate.edu

## EXECUTIVE SUMMARY

The SWITCHBOARD (SWB) Corpus consists of 2430 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.

We have completed the first six months of the SWB project and have released 525 conversations with corrected segmentation, transcriptions, and automatic word alignments. Additionally, there are 275 conversations awaiting release with automatic word alignments. These 800 conversations comprise 41% of the conversations used in the WS'97 partition, and 33% of the entire SWB corpus. We have also performed a major overhaul of the lexicon by removing incorrect or unnecessary entries and making the lexicon case sensitive. Finally, we have created extensive documentation including a statistical analysis of the conversations and a description of the transcription conventions. All such information is on-line and available via the Internet.

In an effort to make the resegmentation process highly efficient, we have developed a segmentation tool that is specifically tailored to the needs of the SWB project. It is written in C++ and uses Tcl-Tk (v8.0) for the user interface. It is highly portable across environments including Windows'95. Our validation staff uses this tool to execute the following tasks:

- **segmentation**: creation of a new segmentation that consists of utterances typically 10 seconds in duration and excised at significant pause boundaries and/or turn boundaries;
- **transcription validation**: correction of the orthographic transcriptions;
- **word alignment**: adjustment of word boundaries produced by a forced alignment that uses the new transcriptions with our best phone-based LVCSR system.

Our cross-validation tests on relatively clean utterances have shown that our validators have an average word error rate (WER) of 2.6% (this number varies dramatically with the convention one uses for scoring). This is a substantial improvement from the 8% WER (measured under similar conditions) present in the current best transcriptions recently released by LDC. After manual word alignments — our final quality control step — the WER is reduced to 1.5%. Our best validators are able to reduce the WER to less than 0.5%. We are currently implementing measures to reduce the average error rate to less than 1%. To place this in perspective, a typical six minute conversation has approximately 1200 words, which implies that the final transcription will have approximately 12 words in error for each conversation.

To further underscore the importance these new transcriptions, we have demonstrated a 1.9% absolute improvement in recognition performance (from 49.7% to 47.8%) simply by training on the new transcriptions. Equally exciting is the fact that recognition error rates on monosyllabic words dropped a similar amount — from 49.6% to 47.7% (and performance on other words dropped from 49.1% to 47.4%). Since monosyllabic words dominate the SWB Corpus, this is a particularly significant result.

## TABLE OF CONTENTS

## ABSTRACT

The SWITCHBOARD Corpus (SWB) is a database of 2430 spontaneous conversations recorded digitally over long distance telephone lines. The conversations average 6 minutes in length, total over 240 hours of data, include over 3 million words of text, and contain 541 unique speakers (300 males and 241 females). Most major American English dialects are contained in this corpus. The word error rate (WER) of the current best reference transcriptions has been measured to be in the range of 10%. Such a high error rate is perceived to be a major stumbling block in the development of improved large vocabulary conversational speech recognition (LVCSR) technology.

It is the goal of this project to resegment the SWB Corpus, to correct the transcriptions such that they have a vanishingly small WER, and to supply relatively accurate word boundary information. Towards this goal we have released 525 conversations with corrected segmentation, transcriptions, and automatic word alignments and have an additional 275 conversations awaiting release with automatic word alignments. These 800 conversations comprise 41% of the conversations used in the WS'97 partition, and 33% of the entire SWB corpus. Additionally, we have developed extensive documentation including a statistical analysis of the conversations, a revamped lexicon, and a detailed transcription conventions document.

We have also demonstrated the benefits of these new transcriptions by conducting a limited recognition experiment using the new data. We have achieved a 1.9% absolute improvement in recognition performance on a standard WS'97 evaluation task by simply training existing Hidden Markov Models (HMM) on about 350 conversations with new transcriptions. Equally exciting is the fact that we obtained an equivalent reduction in WER on monosyllabic words: 49.6% for the original system; 47.7% for the new system. Monosyllabic words are the single most common class of words for SWB and account for about 70% of the errors in a typical recognition system.

## 1. HISTORICAL BACKGROUND

In the early 1990s, DoD and DARPA saw the need for a large amount of data from a variety of speakers to be used for a variety of speech research needs including speech recognition, speaker recognition, and topic spotting. Previous common evaluation tasks, such as the Resource Management (RM) [1] and Air Travel Information System [2] (ATIS) tasks, had been narrow in scope and covered only a few speakers. Texas Instruments was sponsored by DoD in 1990 [3] to collect the SWB Corpus. In 1993, the first LDC release of the corpus occurred. In addition to transcriptions, this release included transcriptions segmented by conversation turn boundaries, and time alignments for each word based on a phone-level supervised recognition.

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. The data was subsequently used for many types of research for which it was never originally intended. Hence, 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 began work to do a final cleanup of the SWB Corpus, and to organize and integrate all existing resources related to the data into this final release.

## 1.1. The Data Collection Paradigm

SWB was the first database collected of its type: two-way conversations collected digitally from the telephone network using a T1 line. In retrospect, a number of issues in this type of data collection have surfaced — most notably a problem involving echo cancellation. In the original SWB data collection, echo cancellation was not always activated because the phone calls were bridged within the SWB data collection platform, and hence appeared as local calls to the network. This resulted in a significant portion of the data having serious echo. As described later, we routinely use echo cancellation during transcription to counteract this. Unfortunately, echo cancellation is not always as effective as we would like.

There are also a variety of real-time problems evident in this corpus. For example, some conversations experience a loss of time synchronization between channels of the data. This causes serious problems for the echo canceller, which assumes a fixed or extremely slowly varying delay between the source signal on one channel, and its echoed version on the other channel. Sometimes the echo appears before the source signal — clearly indicating a loss of data somewhere. Similarly, occasionally data appears to be lost without any corresponding error reports, causing unnatural chops in the audio files on one or both channels. Sometimes the missing data is filled with a run of zero amplitude values. In a related problem, data has been observed that is “out of order” (the latter part of a word comes before the first part of the word) signaling that perhaps buffers have been swapped or overwritten during collection. Finally, some conversations suffer from the introduction of digital noise due to out-of-band signaling. Many of these problems are nicely summarized in an FAQ [4] developed for this project that we maintain on our web site. Its primary purpose, as described later in Section 3.7, is to capture these anomalous cases, and present them to the community for discussion.

## 1.2. A Historical Perspective on the Transcription Problem

SWB, in its entirety, consists of 2430 conversations totaling over 240 hours of two-channel data from 541 unique speakers. The average duration of a conversation is six minutes, as shown in Figure 1. Of the 500 speakers present in the corpus, 50 speakers contributed at least one hour of data to the corpus. A distribution of the amount of data from each speaker is shown in Figure 2. The first half of the database was transcribed by court reporters; the second half by hourly workers employed by TI. Since SWB was one of the first conversational speech corpora of its type, conventions for transcription were extremely controversial, and there was not much of an inventory of prior art [3].

The two main goals of the transcription conventions were consistency and utility in speech and linguistic research. Human readability was also important because it aided in the quality control steps taken after transcriptions were complete. It was decided that conversations would be broken at turn boundaries (points at which the active speaker changed) and use a simple flat ASCII representation for the orthography. Quality control steps included spell checking the transcriptions, checking for misidentification of speakers, and looking for common language or

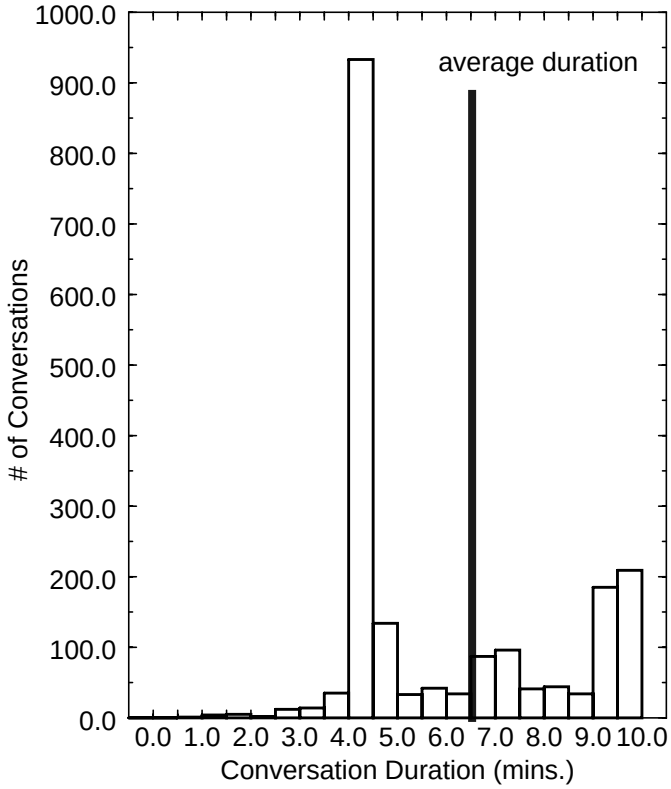


Figure 1. The distribution of the duration of a conversation in SWB shows the mean conversation duration is 6 minutes. The maximum duration was hard-limited to 10 minutes by the data collection system.

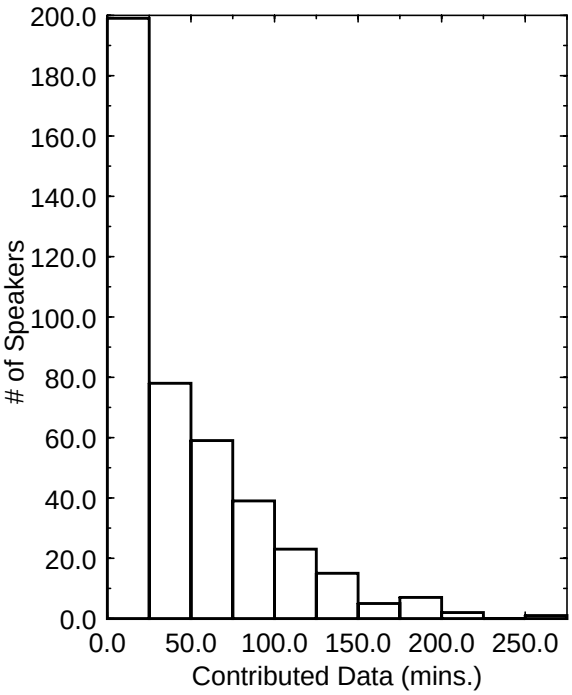


Figure 2. The distribution of the amount of data per speaker in SWB is shown. Subjects were allowed to participate more than once.

spelling errors (its, it's, they're, their, there, etc.). After the transcriptions and quality control steps were complete, time alignments were generated which estimated the beginning time and duration of each word. Finally, a rough check of the time alignments was made by playing samples of each conversation at several places throughout the speech file; errors of over one second usually resulted in reprocessing the data [5].

### 1.3. Segmentation and Its Impact on Technology Development

Initial LVCSR systems had high recognition error rates on SWB — approximately 70% in the early and mid-1990's. The sources of this degraded performance include the lack of a robust language model (which proved to be effective on Wall Street Journal) and poorly calibrated acoustic models (there is a good degree of mismatch between the training and test database when one examines acoustic scores). The difficulties in recognition arise from short words, telephone channel degradation, and disfluent and coarticulated speech. In an effort to reduce error rates, many state-of-the-art systems introduced dynamic pronunciation models [6] and a flexible supervised training procedure [7]. Over the years, WER on various subsets of SWB have fallen to the mid-20% range [8] and in the low 30% on standard evaluations.

However, as performance improvements become less dramatic, and most of the obvious obstacles to performance are overcome, the quality of the training database soon becomes an issue. Casual

reviews of the SWB Corpus as processed by most sites quickly reveals the fact that much of the data is discarded due to the unreliable transcriptions. Pilot studies at WS'97 made it evident that improving the quality of the database through resegmentation and transcription corrections could greatly improve the resultant acoustical models being used for LVCSR experiments. Simply resegmenting the test database resulted in a 2% reduction in WER [9].

In the past, speech segmentation was guided by linguistic or acoustic information metrics. To linguistcally segment data, one places boundaries in natural breaks in speech (between phrases, sentences, turns, etc.). In acoustic segmentation, boundaries are placed in acoustic silence between words. Though both are commonly used, each of these methods has its drawbacks. Both historically have resulted in utterance definitions that have truncated words at the beginning or end of the resulting speech file.

Linguistic segmentation is effective in maintaining clear linguistic context, but it has two important problems. First, if the boundaries are based solely on language rules and not on acoustics, boundaries may be placed between words where there is little or no silence. This will result in word beginnings and ends being cut off which will adversely effect training of acoustic models. Second, linguistcally based boundaries often result in utterances which are too long for experimental recognition systems. Speakers in SWB sometimes carry on monologues of the same thought for 30-60 seconds, but the ideal utterance length for experimentation is closer to 10 seconds (note that common evaluations have often used much shorter utterance definitions).

Segmenting speech based solely on acoustic boundaries also has its advantages. It is a more desirable paradigm in that boundaries are only placed where there is a pause in speech, but this method obscures any inherent linguistic context. Thus, it is of no use when training language models. A major portion of this project involves resegmentation of the data at boundaries that represent a compromise between these two principles: manually placing boundaries where there is acoustical silence, maintaining linguistic context, and regulating the length of the utterances. The net result will be utterance definitions with ample amounts of silence at the beginning and end of the file, and yet contains at the very least a linguistically meaningful phrase or unit. All data is accounted for in our segmentations, so utterance definitions involving larger linguistic units can be easily built from these segmentations.

## 2. SOFTWARE

ISIP began the development of a segmentation tool to facilitate manipulation of SWB conversations. Our interest in this tool stemmed from our desire to continue our research on improving LVCSR performance on monosyllabic words [10,11] through the use of syllable models. Over the first six months of the project, this tool has undergone substantial modifications that reflect our much better understanding of the challenges of segmentation and transcription of SWB. The tool has also pushed through several external design reviews involving potential customers. Their feedback has been invaluable towards making the tool more general and extensible.

An overview of our segmentation tool is given in Figure 3. A screenshot of the Unix version of the tool is shown. This tool is specifically designed to consolidate the tasks of resegmentation, transcription correction, and word alignment review into a single intuitive, yet powerful, package.

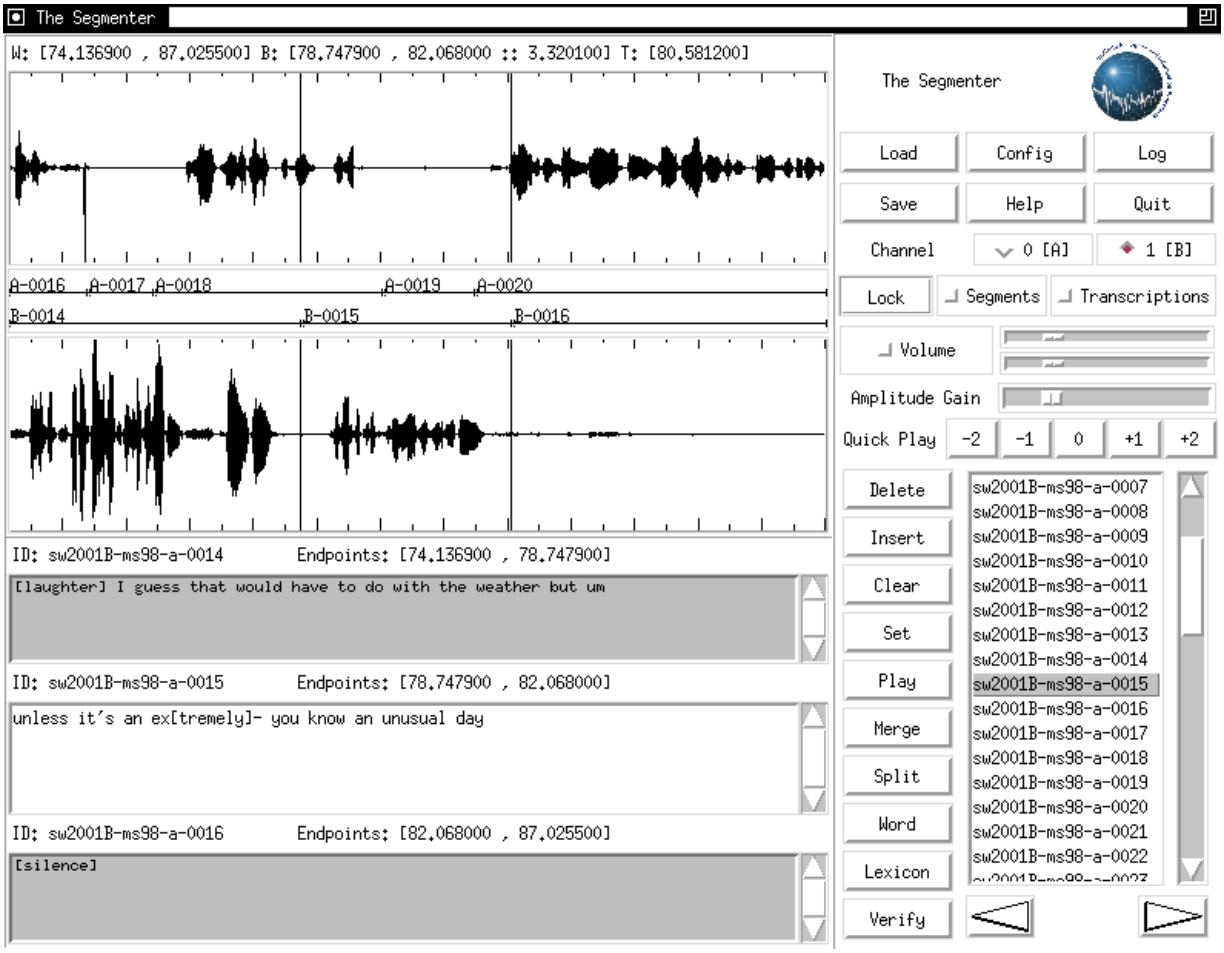


Figure 3. A SWITCHBOARD resegmentation tool that allows for easy manipulation of segmentation and transcriptions of conversations on a per-utterance basis. Transcribers operate at less than 20x real-time with this tool — our best validators can achieve 10x real-time.

It has enabled our validators to efficiently produce highly accurate transcriptions by placing all of the necessary functionality directly at their fingertips. Most functions are executed from accelerator keys — the user rarely needs to take their hands off of the keyboard. A brief introduction to the tool follows.

## 2.1. An Overview of the Segmentation Tool

Our segmentation tool is a graphical, point-and-click interface tool designed to expedite the segmentation/transcription process. This tool, is written entirely in C/C++ interfaced to Tcl/Tk and is designed to be highly portable across platforms (we currently run it on Sun Sparcstations as well as Pentium-based desktops running Solaris; an extension to Windows is available, but does not as yet have a clean audio solution). It also supports numerous audio utilities. The current version of the segmenter is highly customized to be used with the SWB Corpus. However, it is easily extended to other domains (we have demonstrated this with the recent release of a single-channel version of the tool) and is freely available [12] via the Internet.

Our tool has greatly streamlined the segmentation process. Its most fundamental design feature is that all speech data must be accounted for. Silence regions are explicitly marked; no audio data is

ignored in the transcription process. This tool has a short and easy learning curve that results in a short training period for our validators, yet allows them to efficiently alter the utterance boundaries and transcriptions. The display area of the tool provides the validator with instant access to the acoustic waveforms, the audio context for any utterance, as well as the functionality to zoom in and/or play a selected portion of the utterance. An additional word-alignment mode allows the validators to check the transcription accuracy word-by-word at high speeds, thus providing a efficient means of maintaining strict quality control.

The audio tools embedded in our segmentation tool are obviously an important part of the tool. Each channel of the two-channel signal (often mistakenly referred to as a stereo signal) can be reviewed independently, or both channels can be heard simultaneously. Two-channel audio is an integral part of the SWB task, since it allows the transcribers to probe each side of the conversation separately or listen to the full context. This, coupled with the echo cancellation of data, allows transcribers to fix many of the swapped channel problems that have plagued SWB. Merging or splitting utterances is as simple as clicking a button. There are features to delete or clear the transcriptions of the current utterance or to insert a new, blank utterance. Transcriptions are easily modified and convenient key strokes make it easy to move between utterances. A listing of the current set of key bindings available in the tool is given in Figure 4. This provides some insight into the flexibility and comprehensiveness of this tool. More information can be found at the tool's web site [12].

## 2.2. An Overview of the Word Alignment Mode

Our original word alignment tool allowed for viewing the boundaries for each word and for listening to each word of a conversation individually. A screenshot of the word alignment tool is shown in Figure 5. The words in a transcription can be played in a continuous audio stream in which short pauses between are automatically inserted. Typically, initial word alignments come from an automated tool, such as an LVCSR system running in supervised recognition mode. In the word alignment review phase, validators can perform a rough check as to whether these alignments are correct, or need adjustment. If the latter, the same tools as used in utterance segmentation are available to adjust the boundaries.

After the pilot phase of the manual word alignment portion of this project began, we realized the need to incorporate the process of transcription corrections and quality control directly into the word alignment tool. For this reason we added buttons to add, remove, or change words in the word alignment tool. This gave a dramatic reduction in the time consumed in the process of correcting transcription errors found during the word alignment phase. However, as detailed in the next section, this type of review quickly fatigues validators, and does not appear to be feasible on a large scale.

We are currently working with the validators to continue development of the word alignment tool in an effort to make the process even more efficient. With the modifications that were made to the transcription and resegmentation tool, validators have been able to perform much more efficiently than we had expected during resegmentation. Word alignments, on the other hand, are currently requiring more work than budgeted. This, not surprisingly, is due to the fact that individual words are hard to distinguish in SWB — particularly when played with no surrounding acoustic context. Hence, validators find it hard to distinguish between a poorly articulated word and an incorrect

**Signal Plot Area:**

- Left mouse button: mark 1st time bracket
- Drag mouse with left button pressed: move 2nd time bracket
- Left mouse button release: mark 2nd time bracket
- Right mouse button: pop-up menu
- Mouse movement: current time

**Control Panel:**

- Middle mouse button help

**Control Pan. List Box:**

- Left mouse button double-click: set current utterance

**Signal Display:**

- Alt-RightArrow: window ahead
- Alt-LeftArrow: window back
- Alt-UpArrow: zoom out
- Alt-DownArrow: zoom in
- Alt-b: zoom in on brackets
- Alt-z: zoom out full

**Audio Play:**

- Alt-m: between bracket marks
- Alt-w: current window data
- Alt-u: current utterance
- Alt-e0: channel 0 btwn marks
- Alt-e1: channel 1 btwn marks
- Alt-e2: both channels
- Alt-f0: channel 0 window
- Alt-f1: channel 1 window
- Alt-f2: both channels window

**Utterance Properties:**

- Alt-t: set time marks on current utterance

**Utterance List:**

- Alt-i: insert a new utterance
- Alt-d: delete the current utt.
- Alt-g: merge the selected utt.
- Alt-j: split the current utt.

**Utt. List Traversal:**

- Alt-n: move to the next utt.
- Alt-p: move to the prev. utt.

**Load, Save and Quit:**

- Alt-l: load configuration
- Alt-c: configure
- Alt-s: save data
- Alt-q: quit

**Miscellaneous:**

- Alt-a: start word alignments
- Alt-h: help
- Alt-o: set bookmark
- Alt-r: mark utterance
- Alt-v: toggle verify mode
- Alt-x: load lexicon

**Word Alignments:**

- Alt-b: previous word
- Alt-f: next word
- Alt-p: previous utterance
- Alt-n: next utterance
- Alt-q: quit word alignments
- Alt-s: save word alignments
- Alt-d: delete current word
- Alt-i: insert new word
- Alt-r: replace current word

Figure 4. An overview of the key bindings supported in the segmentation tool. Key bindings are easily remapped, are designed to reflect common GNU conventions, and are intended to be fairly intuitive.

boundary assignment. As we did for transcription and resegmentation, we are evaluating the word alignment process and will make any necessary modifications to that tool which will increase the productivity of our validators during manual word alignments.

## 2.3. Integrated Project Management Tools

We have spent a great deal of time tailoring this tool to the needs of this project. The process of splitting and merging utterances, which is crucial to resegmentatin of SWB, has been fine-tuned


Figure 5. Word alignment mode in the segmentation tool allows for easy manipulation of word boundaries and for quick transcription modifications.

to maximize validator accuracy and efficiency. Also, validators can log questions about a specific utterance in a log file for review by the project manager. Of course, this can get quickly out of hand for SWB, where much of the data is highly ambiguous, so such features have to be used with great discretion.

At a very early stage of the project, we realized that productivity feedback would be crucial in motivating the validators to improve their performance. Hence, our tool logs in great detail the real-time performance of the validators through the use of a bookmarking feature that time-stamps the log file as each utterance is processed. This information is post-processed to generate a weekly project report that summarizes validator performance. We have found that this feedback is the single-most useful piece of data for encouraging validators to be as productive as possible. It has generated significant cost-savings to the project in that charged hours more accurately correlate with the amount of data generated, and because the real-time rates of the validators tend to drop (with little impact on accuracy) once they know they are being monitored.

One additional feature that was added to this tool during WS’98 was an ability to lock the segmentation or transcriptions so that changes won’t be made accidently during the review of a conversation. This has made the tool much more useful as a general tool for viewing SWB data, and also improved our ability to easily use this tool as a teaching aide. In fact, students at the Summer Workshop on Language Engineering, hosted by the Center for Language and Speech Processing at Johns Hopkins University, used the tool to learn about SWB. Several researchers also used the tool to listen to selected SWB utterances. Their feedback was invaluable in making modifications to the tool to reduce the start-up costs and infrastructure required to run the tool on new data, as well as increase the number of devices for which there is audio support.

## 3. RESEGMENTATION OF SWB

Preparation of the SWB conversations is a multi-stage process consisting of numerous quality control procedures. A detailed flowchart of the procedure we follow is shown in Figure 6. The illustrated process can be broken down into five major steps: data preparation; segmentation; transcription correction; automatic word alignments and manual word alignment review. Each of these tasks is described in detail below, along with some general comments on quality control. Two auxiliary outputs from this process are described in this section also: the SWB FAQ which represents a collection of interesting examples of problematic utterances, and the SWB Progress Report which is used to monitor validator output on a weekly basis.

### 3.1. Data Preparation

We begin our process of resegmenting SWB by removing the transcriptions and audio files from the SWB release titled “Switchboard-1 Telephone Speech Corpus: Release 2 August, 1997”. We are using the following CDs in this project:

- “Switchboard-1 Transcriptions: Intermediate Version” August, 1997
- “Switchboard-1 Telephone Speech Corpus: Release 2" August, 1997

After downloading this data to our systems, we process the NIST data for use with the segmenter with a script called *prepare_data*. This script converts the sphere files to 16-bit linear raw files, separates the “.mrk” files into separate transcription files for each channel, and echo cancels the data. Past attempts to transcribe SWB have not dealt effectively with the echo present in the audio data. This has caused numerous problems with swapped channels in transcriptions and with incorrect transcriptions. To avoid these problems in our data and to provide the validators with the highest possible audio quality, all conversations have been echo cancelled before transcription. This process consists of simply passing the data through ISIP’s standard least mean-square error echo canceller [13,14] which has been optimized for the SWB task (and is currently used by NIST as a standard preprocessing step for conversational telephone speech data). By allowing validators to play each channel of the audio file separately, and providing them with echo cancelled data, we are once and for all eliminating the swapped channel problem that has perennially plagued the SWB Corpus.

After the data is prepared for resegmentation, it is assigned to a validator. Conversation assignments are based on difficulty level. A validator’s weekly assignment will consist of conversations of all difficulty levels so that the most difficult conversations will be distributed equally among the validation staff. Before the assignments are made, a config file is created for each conversation. This is done by using a script called *create_config* which makes a “.cfg” file containing the conversation number and the login of the validator assigned to the conversation.

### 3.2. Segmentation

Resegmentation of the SWB training database is the most important part of our work on this project. At the 1997 Speech Recognition Workshop, similar resegmentation work on the test database resulted in a 2% reduction [6] in word error rate (WER). Resegmentation is a

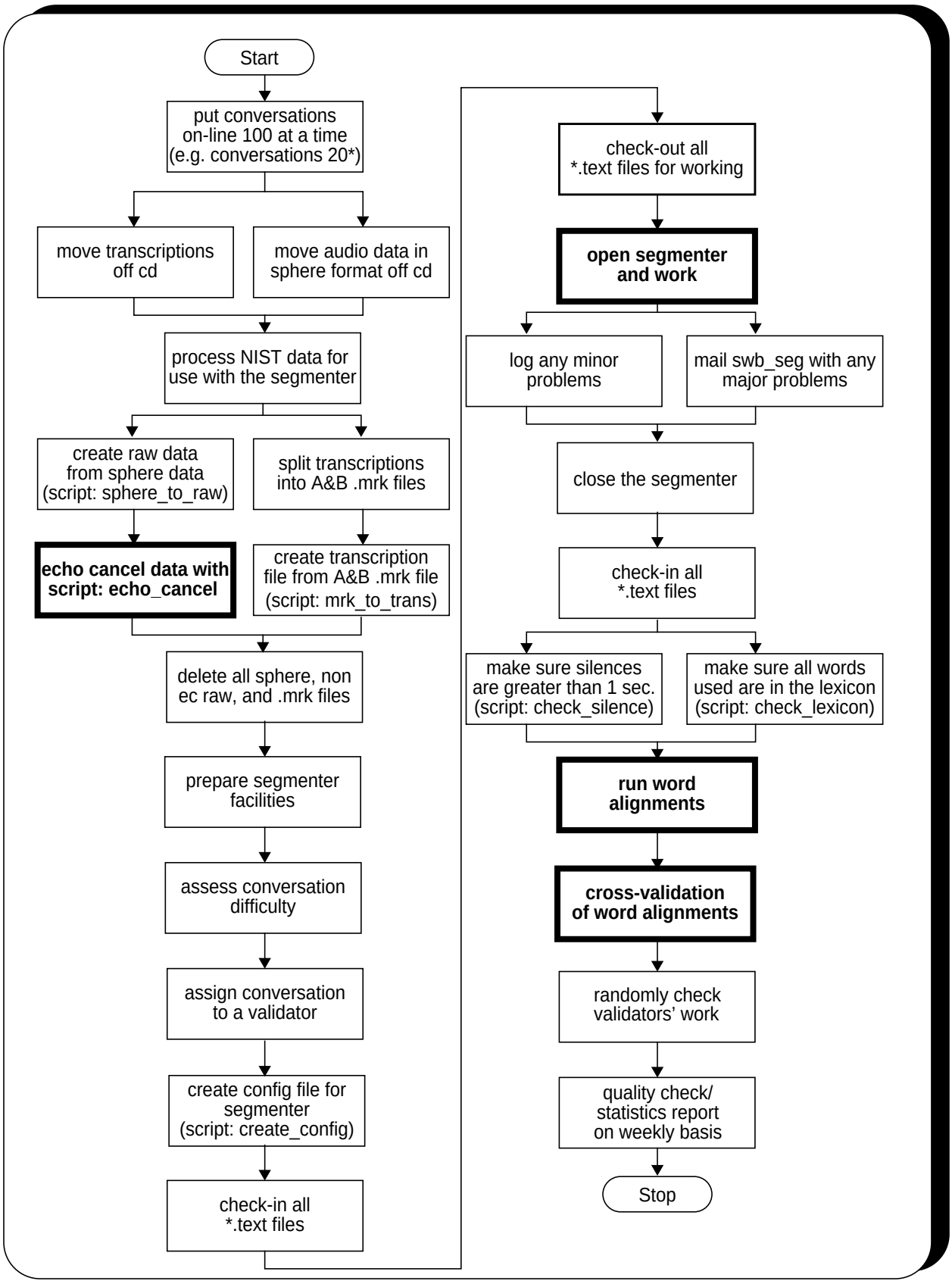


Figure 6. Workflow diagram for SWB resegmentation project.

challenging part of the correction process because a decision must be made on whether to split at natural linguistic boundaries (sentence boundaries, turn boundaries, phrase boundaries, etc.) or to split at acoustical boundaries where there is a pause between speech. Our strategy for resegmentation is as follows:

- Segment at locations where there is clear silence separating each segment (at least 1 second long);
- Segment along phrase, sentence, and/or train-of-thought boundaries.

The first rule is important because it eliminates the problem of truncated words due to segment boundaries falling where there was not enough separation between words. This has a negative effect on training of acoustic models since it diminishes one's ability to accurately model coarticulation effects and it may attribute acoustics to the incorrect word of the coarticulation pair thus training the model with out-of-class data. The second rule is implemented to maintain linguistic context and clarity for speech understanding and language modeling experimentation. We have modified these general guidelines to be specific and easily implementable as possible:

- Set boundaries so that each utterance has a beginning and ending silence buffer of 0.5 seconds
- Utterances should be split to be approximately 10 seconds in length

There are several cases where a speaker carries on a monologue for well over 15 seconds without pausing. Our segmentation rules do not allow for splitting of such a long utterance where there is not an acoustical pause of at least 0.5 seconds. However, utterances over 10 seconds cause problems in recognition and training because they require larger search networks, thus more computational resources. An example of such an utterance is shown in Figure 7. In this case there are two alternatives: allow the utterance to span the 21 seconds or segment at a point such that there is very little silence to pad the resultant utterances. For decisions such as these, we consult experts throughout the speech community on a case-by-case basis.

### 3.3. Transcription Correction

After the boundaries have been properly set, the validators make any necessary corrections to the transcriptions. We have produced a highly detailed list of transcription rules that our validators use to handle transcription of partial words, mispronunciations, and proper nouns. These rules originated from the LDC transcription conventions [5] released with the SWB Corpus. We have made significant changes to the original LDC transcription conventions to ensure the highest level of accuracy and consistency in our transcriptions. A complete description of our modified transcription conventions [15] is maintained on our web site and available for public comment. Most of the conventions described in this document have also been discussed in a mailing list we maintain for this project: *swb@isip.msstate.edu*.

Many of our transcription rules were a by-product of problems pointed out by our validators. Each time that a validator was not able to easily arrive at a transcription by following our conventions, we were compelled to add a rule to help maintain clarity and consistency. Our procedure in such a case is to solicit input from the community to arrive at a consensus, and then inform the validators of the result. Listed below are a few of the more interesting and difficult issues that we have encountered during the first six months of this project:

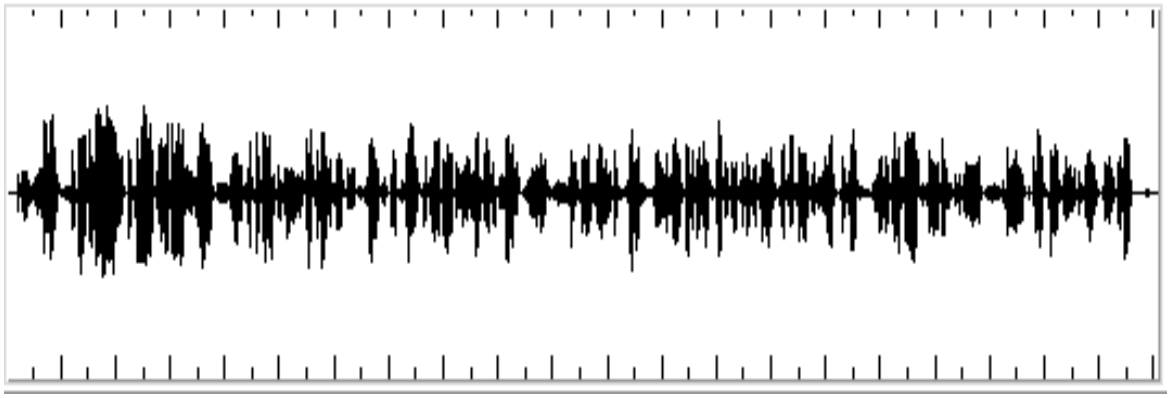

Figure 7. In the above waveform, a speaker provides 21 seconds of continuous speech without an acoustical pause of 0.5 seconds or longer. In such a case, our constraint on the amount of silence padding each utterance must be reduced until a suitable pause can be found. Often this occurs at a point in the data where there is a linguistically meaningful boundary or a string of filler words.

- Title capitalizations: Speakers often refer to titles in their conversations. There is a debate as to how to capitalize these proper nouns. The question was whether we should capitalize each word in a title (example: “Gone With The Wind”) or use standard grammar rules and capitalize the first word, last word, and keep prepositions under five letters lower case (example: “Gone with the Wind”). We decided on the latter option.
- Compound words: It came to our attention that our validators were not being consistent with the transcription of compound words (example: “everyday” vs. “every day”). We decided to transcribe all compound words as one word regardless of context unless there was a definite acoustic pause between the two words.
- Coinages: Speakers often use words in their speech and attribute meaning to these words though they do not occur in the dictionary (example: the person who sells the gun ought to protect themself). In this example, “themself” is not a proper word, but the speaker is using it as if it was. Our convention on these words, called coinages, is to transcribe the word in braces — in this case, “{themself}”.
- Mispronuncations: Occasionally speakers mispronounce a word or say a word they didn’t mean and then correct themselves (example: I blame the splace space program). Here the caller accidently said “splace” and then corrected the mistake by saying “space”. We transcribe such cases with the word they said and the word they meant to say separated with a slash and all enclosed in brackets. The example is corrected as “I blame the [splace/space] space program”.
- Vocalized noise: We have heard several examples of a speaker making a sound that can not be deciphered as a word or partial word and also can not be classified as coughing, breathing, or any of the other usual non-speech noises (example: she was able to pull out of it uh d- w- so cheaply the second time). This speaker uses the “d- w-” as a hesitation sound. Such cases are now transcribed with the tag [vocalized-noise].
- Partial words: Speakers commonly start, but do not finish the acoustics of a word (this is known as a false start) (example: if the speaker began the word “space” but only said “spa-”). Our convention for these cases is to transcribe the part of the word that was said, and enclose the rest of the word in brackets followed or preceded by a dash to keep the context of the word. In this example: “spa[ce]-”.
- Laughter words: The original LDC transcription conventions transcribed laughter alone, but there was no convention for transcribing the act of a person speaking while simultaneously laughing. This occurs quite often so we made the rule to annotate this phenomenon by transcribing laughter and the word spoken separated by a hyphen and all enclosed in brackets. An example is “[laughter-yes]”.

These and many other transcription issues can be found on our regularly updated SWB FAQ [4]. The biggest challenge in transcribing SWB is the transcription of words that are mumbled, distorted, or spoken too quickly by the caller. Even after listening to the words dozens of times and drawing from as much context as possible, there are still times where we must make what amounts to an educated guess. These problems result in most of our final word errors. It could certainly be debated that these sorts of words are of no use for training acoustic models, regardless and, in fact, may be a detriment to the model. However, it is our practice to do our best to transcribe all speech in the database.

### 3.4. Automatic Word Alignments

The process of generating automatic word alignments is rather straightforward with a few minor exceptions. The new segmentations, transcriptions and the echo cancelled data are used to create a new set of word alignments by performing a supervised training with our best phone-based recognizer. We use a crossword triphone system developed during WS'97 [11] and the HTK training tools to run our forced alignments. Our feature set consists of 12 MFCC's, normalized energy, and their corresponding delta and delta-delta features — 39 in all. The methodology used to generate features using HCopy (HTK's feature generation engine) requires that we add 100 samples of silence to the ends of each utterance before generating the features. This ensures that the number of feature vectors generated is equal to the number of frames of data in the utterance. Also during alignment, we require that the utterances start and end in silence. This is a direct consequence of the segmentation process. A diagram of this process is shown in Figure 5.

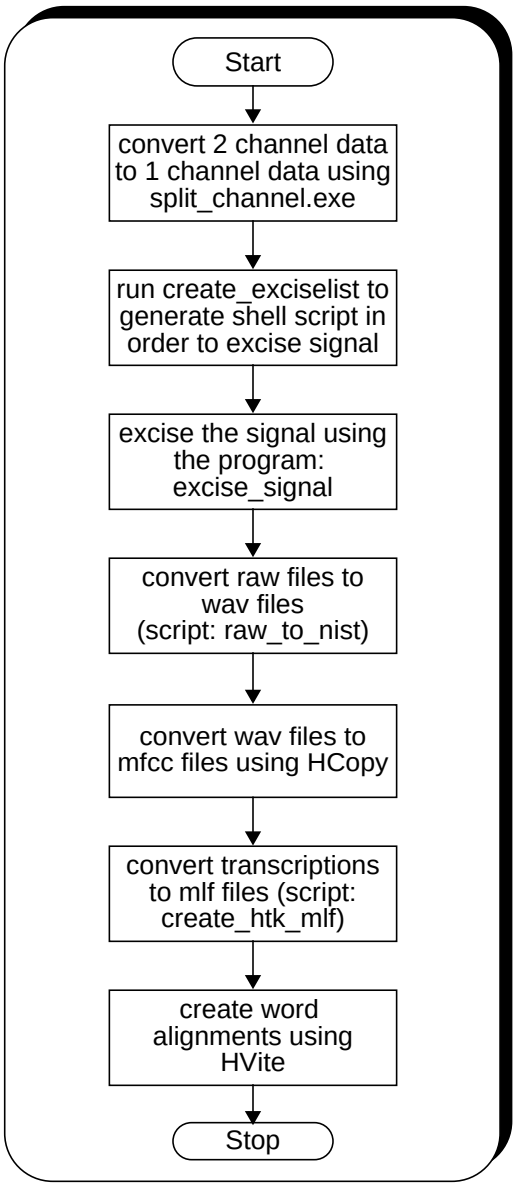


Figure 8. The work flow diagram for generation of automatic word alignments.

### 3.5. Manual Word Alignments

After generation of automatic word alignments is complete, our validators review these word boundaries manually and correct any gross errors. And example screenshot of the word boundaries is shown in Figure 9. This process not only improves the accuracy of the marked word boundaries, but is also our final quality check on the transcriptions. In this phase of the project the validators are looking very closely for any transcription errors and are checking for

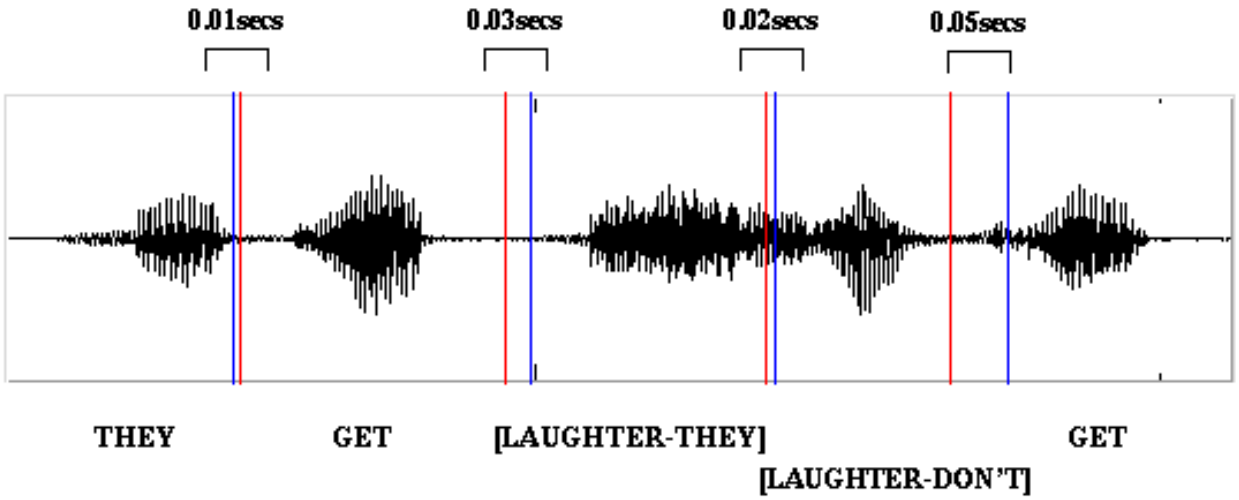


Figure 9. An example of word alignments before and after manual word alignment review is performed. The majority of the problems with the automatic forced alignments center around words bounded by laughter, mouth noises at word boundaries, and partial word pronunciations. In this case, the boundaries determined by automatic alignment are shown in blue; the boundaries after manual word alignment are shown in red. The main change here was that a word boundary was missed for "they don't" which is embedded in laughter. This was corrected with the manual alignments. Also, the last boundary was placed too far into the beginning of the word "get," and was corrected as well.

conformity to all transcription conventions. We find good validators can reduce the transcription word error rates by a factor of 2 or 3 performing this step.

Unfortunately, this part of our SWB project has been the most difficult. We began manual word alignments in April but had a setback when we realized that we weren't properly inserting a silence tag where pauses existed between words. To make these word alignments more accurate, we have recently restarted manual word alignments after changing the process to add silence between words where needed. We are unable to make the recognizer reliably force short silences between words, so we post-processed the recognition output to remove 50 msec or less of silence between words (and simply use a single boundary between words). Validators then review these boundaries correcting gross errors, but do not attempt to precisely adjust word boundaries in situations where there is no discernible silence between the words (to do this would require a spectrogram capability in addition to a large amounts of validator time).

### 3.6. Quality Control

We take several steps to ensure that our released data is of the highest possible quality. After our conversations have been validated, we run three scripts on the transcriptions which check for different kinds of problems. First, we use a script called *check_dictionary* which verifies that each word in the new batch of transcriptions is also present in the SWB dictionary. Words not found in the SWB dictionary are reviewed by the project manager. All acceptable words are assigned pronunciations. This list is further reviewed by two Ph.D. students who correct any errors, and then the words are added to the dictionary. The next quality check uses a script, *check_silence*, to determine the length of silence-only utterances in the transcription files, flagging those that are less than one second long — our standard for minimum silence length. Finally, we run a script

called *check_bounds* which ensures that the start time of every utterance or word is equal to the end time of the previous utterance or word. It also makes sure that the end time of the last utterance or word is equal to the size of the file up to six significant digits. Any flagged errors from these two scripts are corrected in the transcriptions before generating automatic word alignments. After we have generated the automatic word alignment files, we run the script *confirm_word_files* to check for any errors in the word alignments. This script makes sure that the begin and end times of each word file match the begin and end times of the corresponding utterance in the transcription file, checks that all words in the word file match the words in the transcription file, and flags any utterances that do not have corresponding word alignments in the word alignment file. After correcting any errors flagged by this script, the conversations are ready to be released with automatic word alignments and ready to be given to our validators for manual word alignments.

In addition to quality checks of our released data, we conduct blind cross-validation tests to determine the accuracy of each validator and consistency amongst the validators. These comparisons are done using the standard NIST speech recognition scoring package featuring sclite. All errors due to differences between ISIP's transcription conventions and the original LDC conventions are disregarded. Any ambiguous differences in transcriptions such as marking soft breath noise or slight differences in the splitting of partial words are not included in our validators' computed WERs. Thus the results reflect errors that would adversely effect the training of models and other experimentation.

### 3.7. The SWB FAQ

An example of the home page for our SWB Frequently Asked Questions [4] (FAQ) is shown in Figure 10. Clicking on one of the utterances reveals the page shown in Figure 11. Users can play the utterance directly within their browser, and enter their comments on the problem in the dialog box. A click on the "Submit" button logs these comments into the database, and makes them available for viewing. Clicking on "View Comments" will display all comments received to date (posting is immediate) on the item.

The general process flow is that items are added to the FAQ as we encounter them in the transcription process. An item is left open for discussion for a short period of time — typically one or two days. At the end of that time, if a consensus is reached, the resolution of the issue is posted to the web page, and our transcription guidelines document is updated accordingly. If this new policy represents a substantive change of our methodology, we must then go back and fold this change into all previously released data (which is, needless to say, time-consuming).

### 3.8. The SWB Progress Report

An example of our weekly progress report is shown in Figure 12. The most important part of this report is the first block titled "Staffing." Here, we report on validator productivity. The information presented here is generated automatically by scripts that post-processed the log files generated during validation. This is made possible by the bookmarking feature previously described. We maintain detailed logs tracking which validators processed a particular conversation, and manage most of this data using a revision control system (RCS). Such a paper trail is important when tracking errors and diagnosing validator performance problems.

SWITCHBOARD Transcription FAQ

Open for discussion:

Example 034: (09/17/98) [breath-word] or [noise-word]?
Example 033: (09/17/98) Speaker holds the floor with hesitations
Example 032: (09/07/98) I don't understand what makes SWITCHBOARD so hard
Example 031: (08/12/98) "rogo"
• • •

Previously discussed:

Example 021: (06/01/98) Compound words
Example 019: (06/01/98) Mispronunciation or alternate form
Example 016: (06/01/98) "gonna" "wanna" "sorta" "kinda" etc.
• • •

Figure 10. An example of the information contained on the front page of the SWB Transcription FAQ.

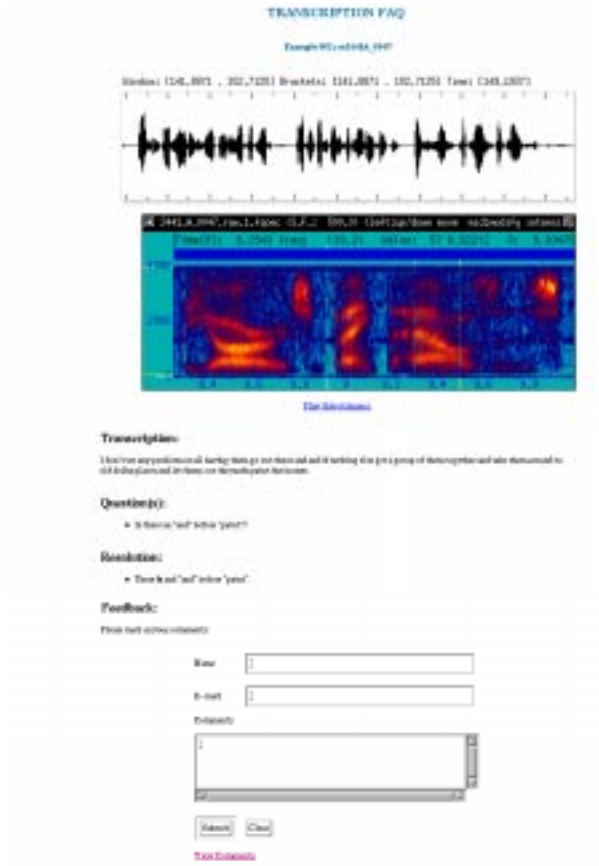


Figure 11. An example of an item available for comments on the SWB FAQ page. Users can listen to the utterance, view the spectrogram (generated off-line), submit comments, and view all existing comments. We hope that by involving the community at this level of the project, we can avoid serious problems with transcription conventions at the end of the project. Unfortunately, participation in the FAQ by external researchers has been low thus far.

SWB Progress Report for Week 8/10/98 - 8/14/98

A. Staffing

| Name | Hours (hrs/hrs) | Trans. (hrs.) | Word (hrs.) | xRT Rate | Hours Logged |
|---|---|---|---|---|---|
| DC | 30/30 | 2.24 | --- | 12.15 | 27.22 |
| Tasha | 32/30 | 1.58 | 0.27 | 13.52 | 32.15 |
| Winfield | 20/20 | 1.08 | --- | 17.45 | 18.82 |

Explanations:

- Tasha worked extra this week to make up for missed hours last week. I don't know how she got her RT rate so fast all of a sudden, but I'm going to look at her accuracy to make sure she isn't letting that slip. If not, this is great!
- I don't know why DC's hours are off, but I will talk to him about it Monday morning and get him to make up the time this week.

B. Production

- DC (27 validated):
  4151 4642 4644 4649 4679 4693 4695 4696 4703 4707 4710 4721 4733
  4734 4735 4736 4745 4788 4792 4799 4801 4802 4805 4812 4821 4822 4826
- Tasha (19 validated):
  4350 4352 4356 4361 4363 4364 4366 4382 4400 4433 4440 4445 4448
  4467 4519 4523 4526 4555 4874
- Winfield (13 validated)
  4171 4174 4175 4177 4181 4182 4315 4317 4318 4323 4324 4325 4326

C. Released Data

- The next 149 conversations are ready and will be released first thing Monday morning. This makes a total release of 525 conversations.

D. Dictionary Development

- Normal additions were made to the dictionary for the new release of 149 conversations.

E. Accomplishments

- Finally completed reviewing and correcting the next 150 conversations.
- Worked on the SWB status report.
- Updated the FAQ.

F. Plans

- Add about 5 more items to the FAQ.
- Find one more validator.
- Make necessary corrections to the SWB status report.
- Prepare for upcoming orientation and site visits.
- Start catching up on releases (we have 275 waiting).
- Work with Tasha and team leaders on word alignment problem.

G. Other Issues

• • •

Figure 12. An example of the SWB Progress Report that is generated each week during the project.

## 4. SUMMARY OF PROGRESS

All projects involving databases have their ups and downs. Despite a few setbacks, we have made excellent progress in the last six months. As of August 15, we have released 525 conversations with new segmentations, corrected transcriptions, and automatic word alignments. There are an additional 275 validated conversations which are awaiting automatic word alignments before release. These 800 conversations sum to over 41% of the WS'97 data and approximately 33% of the entire SWB database. After significant revisions and improvements to the SWB dictionary, we now have a smaller, case-sensitive dictionary. Removing erroneous entries from the dictionary has improved our accuracy and consistency in transcriptions because misspelled or miscapitalized words can now be found with our scripts that would have passed through checks using the original lexicon. We have assembled a stable workforce of validators, most of whom plan on staying with the project until its completion. Performance metrics have shown that these validators are able to achieve superb accuracy at reasonable real-time rates. We have also created extensive documentation including a statistical analysis of the conversations, and detailed transcription conventions documents. All such information is on-line and available via the Internet. Finally, we have conducted preliminary experiments which show that this project may indeed have important consequences for LVCSR research.

### 4.1. Validator Performance

As seems to always be the case with database work, it was a slow process to recruit a group of validators who were dedicated to turning out a quality product. Validation is grueling work and requires a special kind of person — we believe we have now found a group that fits that description. Our validators are now performing at an average real-time rate of 16-17times real-time (xRT) which is under our budgeted rate of 20x real-time. Our best validators are reaching speeds closer to 10 xRT with no loss of accuracy. The current average WER is 2.68% (computed mostly on cross-validation data collected on the WS'97 dev test and eval portions of SWB), but this is expected to decrease since we are going through personnel changes and more than half of our validators are only recently out of training.

We have conducted a cross-validation experiment in which each worker validated the same conversation (sw3909) and their transcriptions were compared for accuracy and consistency. Each validator's transcriptions were checked against a reference determined upon careful review by the project manager and a panel of Ph.D. students. This was a blind test, so the validators were unaware that they would be scored on this particular conversation. We also compared the original LDC transcriptions to the reference to provide an estimate of the improvement in SWB transcriptions after resegmentations. We discovered that the original transcriptions had a fairly high WER of almost 8%. Our validator with the highest WER of over 6% has left the project, so our current group of validators is averaging a WER of less than 3% before manual word alignments. Two of our three current validators are only recently out of the training period, so we expect these WERs to go down. The results for the cross-validation test are shown in Table 1.

We followed this experiment with an analysis of the types of errors most common in the transcription. We are using this analysis to further solidify our transcription process and to assure the highest quality of data. As shown in Table 2, the primary error modality in the LDC transcriptions is the incorrect transcription of contractions. The LDC conventions did have a

| Validator | WER |
|---|---|
| 1 | 2.27% |
| 2 | 2.40% |
| 3 | 3.37% |
| 4 | 6.18% |
| LDC | 7.86% |

Table 1. Word error rates for each validator's transcriptions before manual word alignments, along with the WER for the original LDC transcriptions for conversation sw3909.

| Type of error | % of total errors |
|---|---|
| transcription of contraction as two words | 36% |
| deletion | 27% |
| addition | 6% |
| substitution | 31% |

Table 2. A breakdown of the error modalities for the original LDC transcriptions of conversation sw3909. Consistency of contractions has been observed to be a major problem.

provision for transcribing contractions when the data justified it, so this is a fair comparison with our transcriptions. Almost as common are the errors of omitting words from the transcriptions or transcribing the wrong words. This is not surprising given the high volume of short monosyllabic words (often truncated) in SWB.

Although we did not include errors based on transcription conventions in the WER of the original transcriptions, ISIP's modifications to the transcription conventions have greatly improved the accuracy and quality of the database. Our conventions provide instructions for transcribing speech with great detail. For example, the original transcription conventions lacked provisions for transcribing laughter during speech and partial word content. Since these two cases occur frequently in SWB, adding rules for transcribing them greatly improved the accuracy of the transcriptions by more closely transcribing the actual speech. Our conventions have been modified with the goal of transcribing all speech in SWB as exactly as possible.

One of our validators did manual word alignments for the transcriptions with the highest error rate and brought the WER down to 1.65%. She was still in training for word alignments when she participated in this experiment, so after a validator is in production mode and we have streamlined the word alignment process, our final WER should be 1% or less which is a tremendous improvement over the original WER of almost 8%.

## 4.2. Summary of SWB Statistics

We have composed several documents that constitute a comprehensive statistical guide to SWB [16,17]. These can be found in the section titled ATTACHMENTS appearing at the end of this document. These documents organize SWB in several different ways to make understanding this large database easier. For example, these documents were used extensively at WS'98 to construct various subsets of SWB, including a test set definition suitable for speaker adaptation research and evaluation. These documents are briefly described in Table 3.

The data included in these files represents our attempt to merge the most current versions of this information within the community. The bulk of the data came from LDC and the JHU summer workshop inventory. Other sources include NIST, CMU, SRI, and BBN. As we proceed through the data, we are updating this information as appropriate.

| Document | Description |
|---|---|
| Topic Statistics | definition of each topic used in SWB conversations |
| Caller Statistics | information about each speaker including speaker identification number, sex, dialect, total number of minutes of speech, and the total number of conversations in SWB and in each WS'97 set |
| Speaker Statistics | statistics for each speaker including total amount of speech data, amount of speech used for WS'97, and total amount of acoustic data for each conversation (categorized by conversations the in which the speaker participated) |
| All Conversations | statistics for each conversation including speaker identification number, topic, and difficulty (as defined originally by TI) |
| Conversation Statistics | detailed statistics for each conversation including speaker number and gender, topic, difficulty, the WS'97 set it belongs to, total acoustic data and transcribed speech, and total acoustic data and transcribed speech used for WS'97 |
| CD Location/Sphere File Sizes | location on LDC CDs and sphere file sizes for each conversation in the SWB database |
| Missing ".mrk" Files | conversations for which ".mrk" and ".txt" files were not released on the LDC CDs |

Table 3. A brief description of each document available in our collection of SWB statistics documents. These documents merge the most recent corrections available for SWB from LDC, and have been generated by pooling a number of sources available within the community.

Another equally interesting and important resource is our SWB Education [18] page.This was originally developed to train students at the WS'98 Summer Workshop. It contains references to publications, data, useful related software, and a historical summary of recognition performance. Further, several important partitions of the SWB Corpus that have been used for published research are included. Also, there are links to lexicons, dictionaries, and other such linguistic resources that are useful for developing word pronunciations and language models.

### 4.3. Preliminary LVCSR Experiments

Of course, the goal of this work is to improve LVCSR performance on SWB. Not surprisingly, monosyllabic words dominate the corpus in terms of word tokens and errors [11]. A natural question to ask is what happens if we update our best Hidden Markov models (HMMs) on the new transcriptions that have been described in this report. There are a number of practical problems that currently prevent us from doing this experiment as thoroughly as we would like. Most notably, we do not have an in-house capability for generating good lattices, so we can't evaluate on the retranscribed dev test and evaluation databases. Instead, we can simply reestimate models and evaluate on the existing test database. In such a scenario, there is a great potential for language model mismatch to unduly influence the results.

In this section, we describe a very preliminary experiment in model reestimation that we believe demonstrates the potential of a corrected corpus. We have adapted existing acoustic models and trained these models with a training set of 376 resegmented conversation (about 20 hours of

speech including silence). The data was evaluated on existing WS'97 lattices so that we could get a quick preview of the potential improvements in WER. Four reestimation passes were made. Laughter was used to update the baseline silence model, and words containing laughter were substituted with their baseform.

The adapted models achieved a 1.9% absolute improvement in WER over the baseline system. The results of this experiment are shown in Table 4. This is a significant improvement that rivals the type of improvements one expects from algorithm advances. We further analyzed this result by sorting errors based on whether or not the error involved a monosyllabic word. We see, not surprisingly, that the 1.9% WER improvement also was observed for monosyllabic words. In other words, the new transcriptions have in fact helped improve the overall performance of the system on monosyllabic words. Equally encouraging is the fact that performance on non-monosyllabic words also showed similar improvements (the only negative point is that insertions rose slightly). Our significant improvement in recognition of monosyllabic words on this limited experiment is a preview of things to come on the entire database. We expect the gain due to the new transcriptions will be comparable to that achieved over one to two years of algorithm research (based on results cited in the common evaluations).

## 5. PLANS AND ISSUES

As of August 15, we have released 525 conversations with new segmentations, transcriptions, and automatic word alignments. In addition to our released conversations, we have 275 conversations that are being prepared for the next release. These 800 conversations comprise 41% of the WS'97 database and 33% of the entire SWB corpus. We also have a vastly improved, case sensitive dictionary to accompany this release.

We are currently on schedule with our planned completion of the project in December 1999. Although we are behind on manual word alignments, our unexpected increase in validation speed has made up for this set back. We will have all 2430 conversations validated with automatic word alignments by early fall of 1999 and will spend the rest of the year completing manual word alignments for the entire database Our most important obstacle in keeping on schedule with this project is keeping a full validation staff at all times and to maintain 80 hours of validation each week. Two of our current validators plan to continue work with the project until its completion, so we should only have turnaround in two validator positions. A detailed timeline of our progress for the next 18 months is shown in Figure 13.

| Error Rate | ISIP | WS'97 |
|---|---|---|
| combined: | 47.9% | 49.8% |
| monosyllabic | 47.7% | 49.6% |
| other | 47.4% | 47.7% |
| correct words | 55.8% | 53.1% |
| substitutions | 31.6% | 32.2% |
| deletions | 12.6% | 14.8% |
| insertions | 3.7% | 2.9% |

Table 4. The results of a preliminary experiment in which HMMs were reestimated on the new transcriptions, and evaluated on WS'97 dev test. Note that almost a 2% decrease in WER was achieved.

We had a few setbacks in the early stages of this project, but we have overcome those to make excellent progress in the last six months. In the first three months of the project, we had a difficult time finding and keeping four validators who could produce the level of quality in their

work that this job requires. But, we eventually found a good staff and were able to get into production mode beginning in April and throughout the summer.

As we progressed through the project, a few problems with the lexicon came up that we did not foresee. There were many errors in the original lexicon that needed to be corrected. We stripped the original lexicon of incorrect, duplicate, or unnecessary entries. We also decided to make the dictionary and lexicon case-sensitive. This involved removing all proper nouns from the dictionary and correcting any capitalization mistakes in the transcriptions. Although this work with the dictionary was very time consuming, it greatly improved our accuracy. With the new case sensitive dictionary, capitalization errors can be found in the transcriptions before they are released.

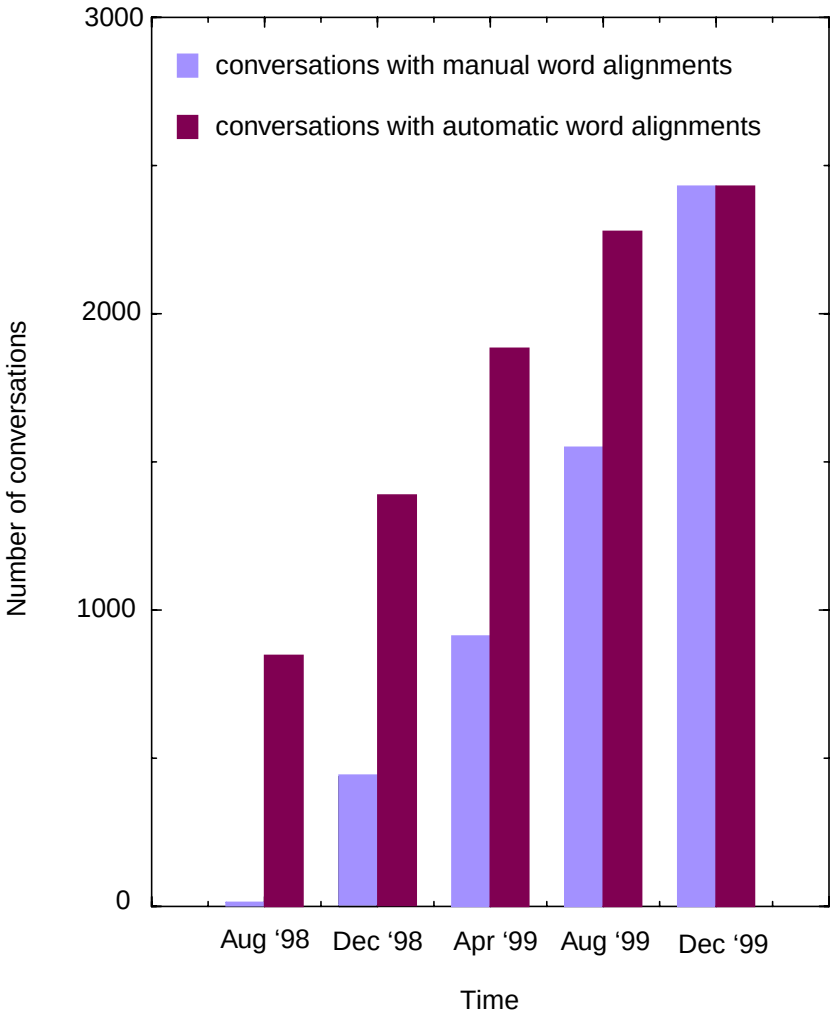


Figure 13. Timeline for the remainder of the SWB resegmentation project. Our anticipated completion date is December 1999.

Another big problem we encountered was manual word alignments. We began manual word alignments in April on the WS'97 dev test set and eval set. After further evaluation and thought about this process, we realized that there was a problem with our generation of automatic word alignments. We were not adding a silence tag where there was silence between words. Unfortunately, we decided that it would be best to throw out the word alignments that we had done to that point and restart them all. After changing the process of generating word alignments so that silence would be entered between words, we prepared to restart all word alignments. This was at the same time we were making the dictionary case-sensitive, so we waited until we completed that revision so that word alignments would also be case sensitive. After six months of having our workers devote all of their time to validation, we will now have them divide their time between validation and manual word alignments.

Our procedure for manual word alignments clearly needs optimization. In the present mode, in which validators listen to an utterance word-by-word, validators find this work very hard and

seem to quickly fatigue. Further, we have recently discovered that on difficult conversations, listening to the utterance word-by-word is not as effective as we thought. Over the next three to four months of this project, we will be revising our word alignment procedure to make it more palatable to the validators, and to improve their accuracy and efficiency. We must arrive at a better procedure if we are to maintain the timeline above.

## 6. ACKNOWLEDGEMENTS

There have been many who have contributed greatly to the SWB Corpus and our resegmentation efforts. To all who have given comments, suggestions and encouragement during the first months of this project, we extend our deepest appreciation. In particular we wish to thank the LDC for supplying us with copies of the SWB CDs and for helping us with many transcription and database-related issues. Secondly, we want to express our continued appreciation to Dr. Jack Godfrey for his continued support in all things related to the SWB Corpus, linguistics, and data collection.

## 7. REFERENCES

[1] P.J. Price, W.M. Fisher, J. Bernstein, D.S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 651-654, New York, New York, USA, April 1988.

[2] C.T. Hemphill, J.J. Godfrey, and G.R. Doddington, "The ATIS Spoken Language Systems Pilot Corpus," *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 96-101, Pittsburgh, Pennsylvania, USA, June 1990.

[3] 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.

[4] 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.

[5] 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.

[6] B. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation Modelling for Conversational Speech Recognition: A Status Report from WS'97," *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 26-33, Santa Barbara, California, USA, December 1997.

[7] M. Finke and A. Waibel, "Flexible Transcription Alignment," *Proceedings of the IEEE*

*Automatic Speech Recognition and Understanding Workshop*, pp. 33-40, Santa Barbara, California, USA, December 1997.

[8] A. Martin, J. Fiscus, W. Fisher, D. Pallet, and M. Przybocki, "System Descriptions and Performance Summary," presented at the Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation, Baltimore, Maryland, USA, May 1997.

[9] B. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation Modelling," presented at the 1997 Summer Workshop on Innovative Techniques for Large Vocabulary Conversational Speech Recognition, the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, USA, August 1997.

[10] A. Ganapathiraju, J. Hamaker, and J. Picone, "Syllable-Based Large Vocabulary Continuous Speech Recognition," submitted to the IEEE Transactions on Speech and Audio Processing, December 1997.

[11] A. Ganapathiraju, V. Goel, J. Picone, A. Corrada, G. Doddington, K. Kirchoff, M. Ordowski, and B. Wheatley, "Syllable — A Promising Recognition Unit for LVCSR," *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 207-214, Santa Barbara, California, USA, December 1997

[12] 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.

[13] J. Picone, M.A. Johnson, and W.T. Hartwell, "Enhancing Speech Recognition Performance with Echo Cancellation," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 529-532, New York, New York, USA, April 1988.

[14] A. Ganapathiraju and J. Picone, "A Least-Mean Square Error (LMS) Echo Canceller," *http://www.isip.msstate.edu/resources/technology/software/1996/fir_echo_canceller,* Institute for Signal and Information Processing, Mississippi State University, December 1996.

[15] 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.

[16] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of SWITCHBOARD," to be presented at the *International Conference on Spoken Language Processing*, Sydney, Australia, November 1998.

[17] 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.

[18] 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.

## ATTACHMENTS

The documents listed below are available on the ISIP web site at the URL shown.

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

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

[3] "A Statistical Guide to SWITCHBOARD: Speaker Statistics", *http://www.isip. msstate.edu/resources/technology/projects/current/switchboard/doc/statistics/ speaker_ stats.text*, Institute for Signal and Information Processing, Mississippi State University, August 1998.

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

[5] "A Statistical Guide to Switchboard: CD Location and Sphere File Sizes," *http://www.isip.msstate.edu/resources/technology/projects/current/switchboard/doc/statistics/cd_location.text*, Institute for Signal and Information Processing, Mississippi State University, August 1998.

[6] "A Statistical Guide to SWITCHBOARD: Missing MRK and TXT Files," *http://www.isip.msstate.edu/resources/technology/projects/current/switchboard/doc/statistics/missing_mrk_files.text*, Institute for Signal and Information Processing, Mississippi State University, August 1998.

[7] "A Statistical Guide to Switchboard: All Conversations," *http://www.isip.msstate.edu/resources/technology/projects/current/switchboard/doc/ statistics/conv_all.text*, Institute for Signal and Information Processing, Mississippi State University, August 1998.