# SWITCHBOARD

By

Vishwanath Mantha

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Electrical Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

August 2000

# SWITCHBOARD

By

Vishwanath Mantha

Approved:

| | |
|---|---|
| Joseph Picone<br>Associate Professor of Electrical and Computer Engineering<br>(Director of Thesis) | Nicholas Younan<br>Professor of Electrical and Computer Engineering<br>(Committee Member) |
| Eric Hansen<br>Assistant Professor of Computer Science and Engineering<br>(Committee Member) | Nicholas Younan<br>Graduate Coordinator of Electrical Engineering in the Department of Electrical and Computer Engineering |
| G. Marshall Molen<br>Department Head of the Department of Electrical and Computer Engineering | A. Wayne Bennett<br>Dean of the College of Engineering |
| Richard D. Koshel<br>Dean of the Graduate School | |

Name: Vishwanath Mantha

Date of Degree: August 13, 2000

Institution: Mississippi State University

Major Field: Electrical Engineering

Major Professor: Dr. Joseph Picone

Title of Study: SWITCHBOARD

Pages in Study: 58

Candidate for Degree of Master of Science

# DEDICATION

I would like to dedicate this to work to XYZ.

# ACKNOWLEDGMENTS

I would lie to thank ABC.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER I

# INTRODUCTION

Speech recognition has developed considerably over the past few decades and is moving closer to the mainstream of society. It is already finding its niche in the medical and legal communities, where specialized vocabularies are used. Speech recognition systems are able to achieve a very high accuracy rate (in the 90 -95% range) for carefully articulated dictation tasks with a fairly general vocabulary [1].

Yet, speech recognition is by no means a solved problem. Performance of contemporary state-of-the-art systems on various tasks [2] is shown in Figure 1. It can be noted that the present challenge lies in recognition of spontaneous conversational speech. Researchers have been tackling this problem for the last couple of decades and have made considerable progress. State-of-the-art Large Vocabulary Continuous Speech Recognition (LVCSR) systems are now able to perform with accuracy levels in the 70-75% range [3]. This, compared to a an accuracy of 30-40% in the last decade, is a very positive sign.

LVCSR research is mainly built upon statistical modeling techniques such as Hidden Markov Models. The implementation of these statistical techniques coupled with natural language science have given a tremendous thrust to LVCSR research. A timeline of improvement in technology of these systems on various tasks is given in Figure 11. Yet, these systems are still very specialized and not accessible to the common man. It should be

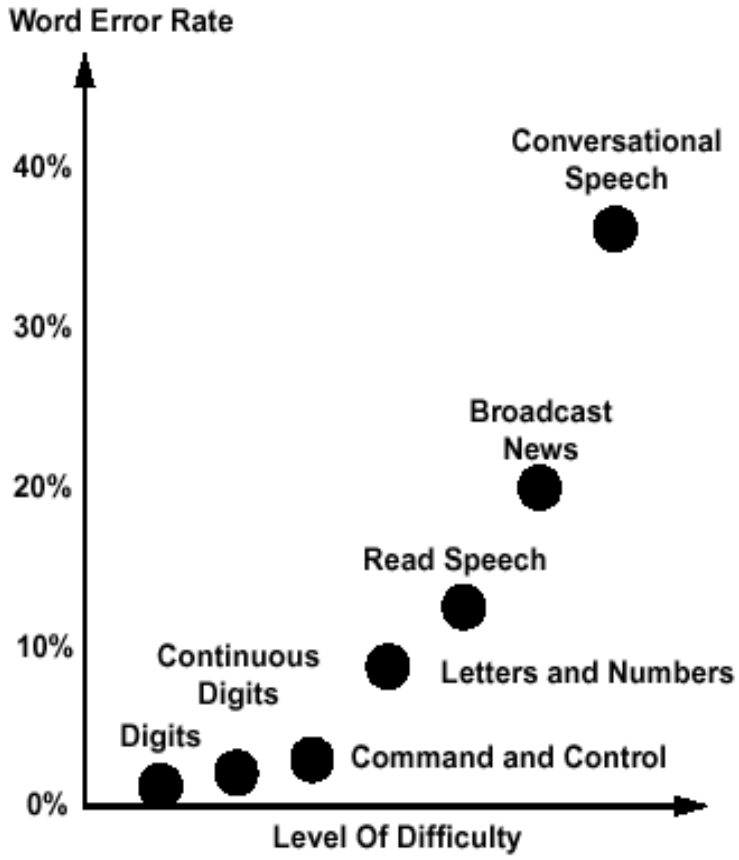


Figure 1. Performance of recognition systems on different tasks

noted that the performance of systems degrades by as much as 2 to 4 times when the systems are integrated into applications for day-to-day usage. The challenge also lies in deploying LVCSR systems in an easy-to-use environment. We still have a long way to go before we can throw away the omnipresent mouse and keyboard and communicate with computers using conversational speech.

## Motivation

LVCSR systems have been the prime focus for almost all the major speech labs in the US for the last couple of decades. Present day LVCSR systems are built using complex statistical modeling techniques. They also incorporate knowledge from other fields of

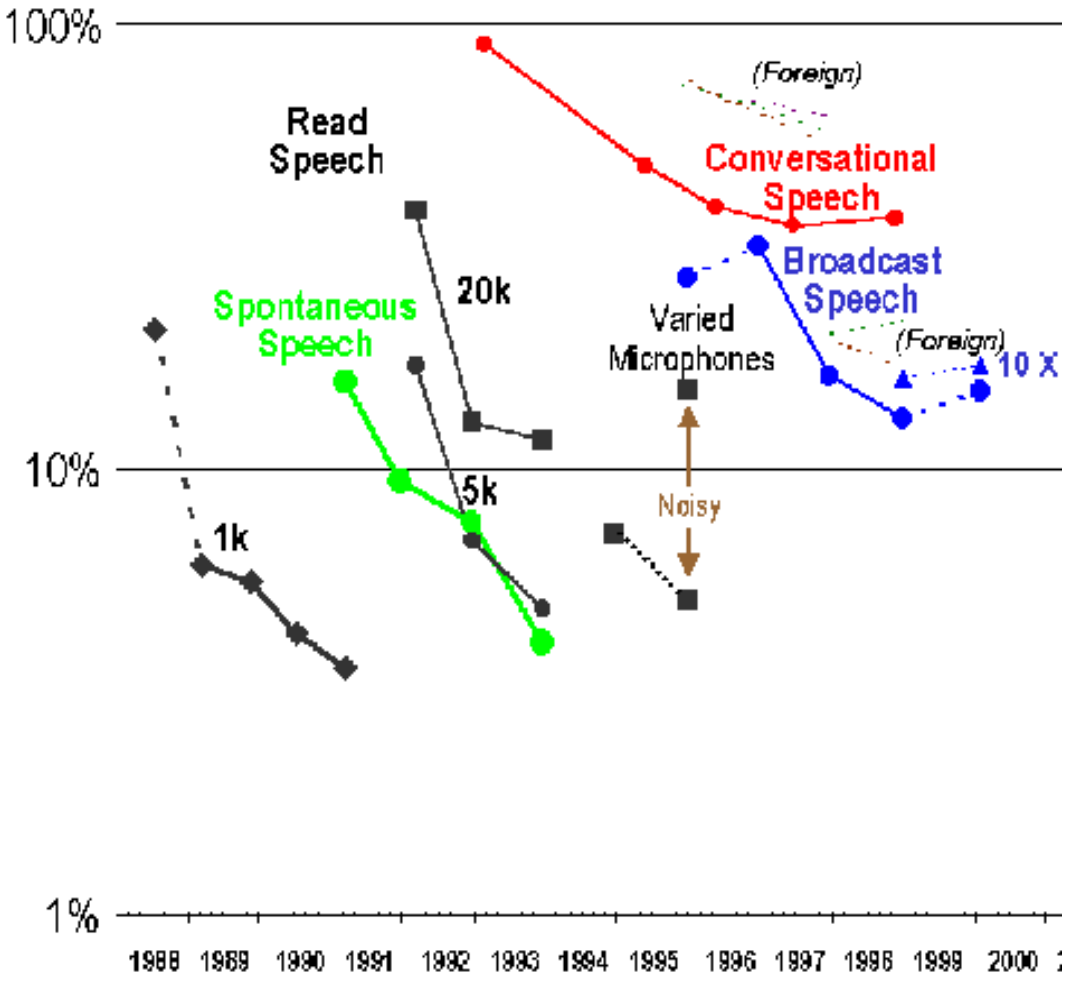


Figure 2. A timeline of speech recognition technology

science such as natural language processing, pattern recognition, artificial intelligence and machine learning [4]. These techniques are an excellent example of data-driven approaches. The training of these systems is done in a supervised learning mode, where the systems learn discriminatory information from the training data. This information is later used in the recognition phase. Hence, it is vital that LVCSR systems be trained on

appropriate data. The use of relevant training data, or the lack thereof, can be the most important factor affecting recognition performance.

Research on identifying appropriate databases is also being carried out in the major speech labs. Government funding agencies have also played an important role in pushing for excellence in this field. The Defense Advanced Research Projects Agency (DARPA) and the Department of Defense (DoD) have been involved in this for the last couple of decades. They have also developed evaluation paradigms for LVCSR systems [5].

Prior to 1990, databases such as Resource Management (RM) [6] and Air Travel Information Services (ATIS) [7] were widely used for evaluating LVCSR systems. These were very narrow in scope and did not cover a large number of speakers. In the early 1990's, DARPA and DoD began to realize the need for a large amount of data from a variety of speakers to address their requirements. After a detailed analysis, it was decided that the recording of telephone conversations would provide the kind of speech data needed. Texas Instruments was sponsored by DoD [8] to collect a telephone database that would cover a large number of speakers and dialects of American English. This was how the SWITCHBOARD (SWB) Corpus came into being. SWB was also intended to facilitate research in speaker identification, topic spotting etc. Hence it was necessary that SWB be a much larger database than the ones mentioned earlier. The original SWB Corpus consists of 2438 conversations spanning 541 speakers and amounted to 240 hours of speech. Since then, other forms of this database were also collected to address newer needs (such as cellular phone conversations).

## Organization of thesis

The goal of this thesis is to study the effect of segmentations and transcriptions of the SWITCHBOARD speech database on LVCSR performance. In Chapter II, a brief description of speech recognition algorithms will be given followed by a summary of the first SWB release. This will be followed by a section that describes the advantages and drawbacks of the original SWB segmentations and transcriptions and will highlight the need for better segmentations and transcriptions. Chapter III will focus on the new conventions developed for resegmenting and retranscribing the SWB Corpus. Chapter IV will analyze the performance of humans on transcription tasks and will also define a quality control process to improve the segmentations and transcriptions. Experimental evidence on the improvements in the new transcriptions will also be reported. In Chapter V, we analyze the new transcriptions and compare their content with the previous generation of transcriptions. The advantages of using the new transcriptions will be discussed. We will conclude this thesis with a chapter summarizing our findings and suggesting future work.

# CHAPTER II

# HISTORICAL PERSPECTIVE

Since the advent of computers, researchers have been trying to build a truly generic human-computer interface. Speech recognition is a direct outcome of this. The initial focus was on simple tasks such as building speech-enabled typewriters. This gradually led to development of elementary speech recognition systems. Organizations at Bell Labs, MIT and other top research labs began spending considerable resources to solve this problem. The gains from this research further fueled the need for LVCSR systems. The technology has also evolved from using dynamic time-warping techniques [4] to that of statistical modeling using HMMs. The next section will give a brief overview of present day LVCSR technology.

## System Architecture

As mentioned earlier, present day LVCSR systems use statistical modeling techniques and also incorporate knowledge about the language being recognized. A typical LVCSR system comprises of the following components:

- Acoustic Front-end
- Acoustic Models
- Language Model
- Search

The block diagram for a LVCSR recognizer is given in Figure 3. The first block deals mainly with the signal processing aspects of speech recognition. The later blocks are involved with statistical aspects of speech recognition. A mathematical formulation of the speech recognizer design is given here. The following subsections will give a brief overview of each of the components of a LVCSR system.

Let $\boldsymbol{A}$ denote the acoustic data (the output of the front-end). $\boldsymbol{A}$ is a sequence of symbols taken from an alphabet $\Im$ :

$$\boldsymbol{A} = a_1, a_2, \ldots, a_m \qquad a_i \subset \Im \tag{1}$$

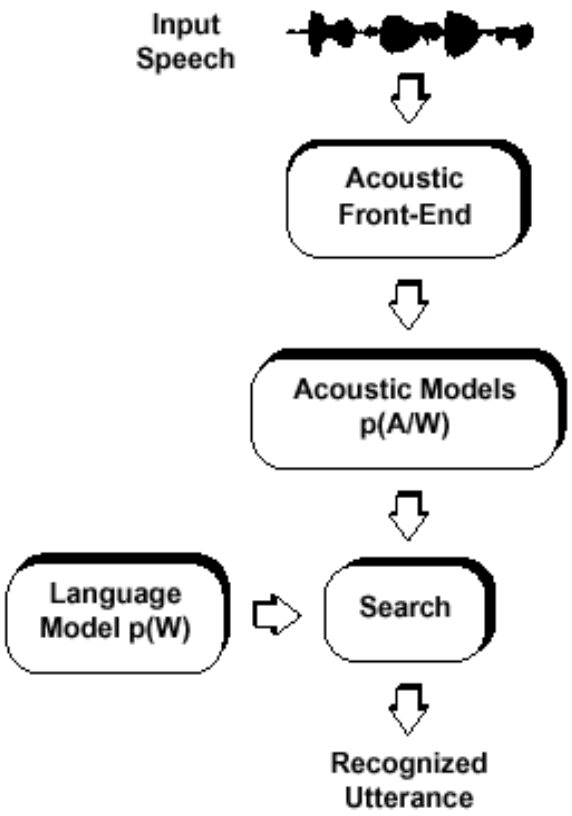


Figure 3. Block diagram of an LVCSR system

The known vocabulary can be formulated as a list of words $w_i$, each belonging to a known vocabulary $\Re$:

$$\boldsymbol{W} = w_1, w_2, \ldots, w_n \qquad w_i \subset \Re \tag{2}$$

If we let $P(\boldsymbol{W}/\boldsymbol{A})$ denote the probability that the words $\boldsymbol{W}$ were spoken, the recognizer should be designed to decide the word $\hat{\boldsymbol{W}}$ which satisfies the following equation:

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\arg max} P(\boldsymbol{W}/\boldsymbol{A}) \tag{3}$$

The right side of the above equation can rewritten using the well known Baye's formula of probability theory. The new equation becomes:

$$P(\boldsymbol{W}/\boldsymbol{A}) = \frac{P(\boldsymbol{W})P(\boldsymbol{A}/\boldsymbol{W})}{P(\boldsymbol{A}))} \tag{4}$$

where $P(\boldsymbol{W})$ is the probability that the series of words $\boldsymbol{W}$ will be uttered, $P(\boldsymbol{A}/\boldsymbol{W})$ is the probability that the acoustic data $\boldsymbol{A}$ will be observed when the speaker says the words $\boldsymbol{W}$ and $P(\boldsymbol{A})$ is the average probability that $\boldsymbol{A}$ will be observed. Since the maximization of (3) is carried out with $\boldsymbol{A}$ staying fixed, it follows from (3) and (4) that the recognizers aim is to find the word string $\hat{\boldsymbol{W}}$ that maximizes the product $P(\boldsymbol{W})P(\boldsymbol{A}/\boldsymbol{W})$:

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\arg max} P(\boldsymbol{W})P(\boldsymbol{A}/\boldsymbol{W}) \tag{5}$$

The acoustic modeling aspect of a speech recognizer deals with the computation of the probability $P(\boldsymbol{A}/\boldsymbol{W})$ and the language modeling aspect of the speech recognizer deals with estimating the probability $P(\boldsymbol{W})$.

## Acoustic Front-end

The purpose of a front-end is to process the speech signal into a series of observation vectors representing events in the probability space [9]. These acoustic vectors used by a network search algorithm to find the most probable sequence of events to hypothesize the textual content of the audio signal. This implies that the front-end is responsible for the type of acoustic data $\boldsymbol{A}$ being fed to a recognizer.

In order for the front-end to model useful observation vectors for LVCSR purposes, it must extract those features from the speech signal that are relatively insensitive to the talker and channel variability which is unrelated to the message content [10]. Most of the contemporary acoustic front-ends are composed of standard signal processing techniques, such as digital filter banks, linear predictive coding and homomorphic analysis. The most common observation vectors used are the Mel Frequency Cepstral Coefficients (MFCCs) [11]. A typical front-end architecture is shown in Figure 4. An in-depth explanation of each of the algorithms is give in [12].

## Acoustic Modeling

To be able to compute the probability $P(\boldsymbol{A}/\boldsymbol{W})$, statistical models need to be built for the speaker's interaction with the system. These statistical models need to model information such as the speaker's pronunciation, ambient noise content etc. The most

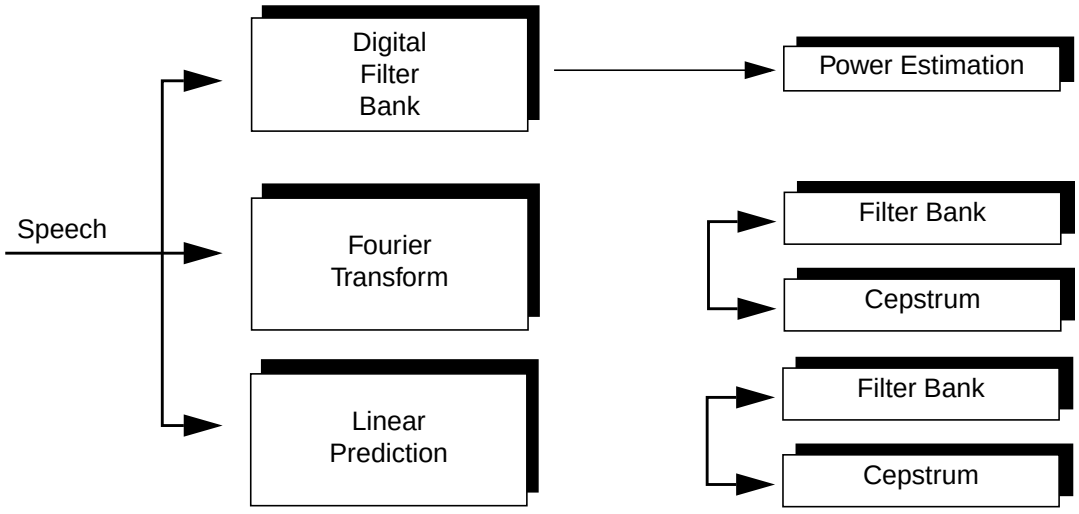


Figure 4. A typical front-end architecture

common acoustic models used in contemporary systems are the Hidden Markov Models (HMMs). HMMs provide an elegant framework to represent the time sequential nature of the speech signal as well as the variability in different sounds.

HMMs are finite state machines in their most basic form. Unlike regular finite state machines, they also allow states to emit symbols with a probability distribution [13]. A HMM can be uniquely characterized by the following parameters:

- $N$—the number of states
- The state-transition probability distribution $\underline{A} = \{a_{ij}\}$
- The output probability distribution $\underline{B} = \{b_j(\boldsymbol{o})\}$, where $\boldsymbol{o}$ is the input observation vector

The output probability distribution is often modeled as a multivariate Gaussian distribution and can be written as:

$$b_j(\boldsymbol{o}_t) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} \exp\left(-\frac{1}{2}(\boldsymbol{o}_t - \mu_j)'\Sigma_j^{-1}(\boldsymbol{o}_t - \mu_j)\right) \quad (6)$$

where $\boldsymbol{o}_t$ is the observation vector at time $t$ and the subscript $j$ indicates that the Gaussian under consideration belongs to the $jth$ state. A simple five state continuous density HMM is shown in Figure 5. The parameters of the HMM are estimates using standard training procedures like Viterbi training or Baum-Welch training. A detailed analysis of these techniques is given in [14].

Language Modeling

Language modeling deals with the estimation of $P(\boldsymbol{W})$, i.e, the a priori probability that the speaker wishes to say the word $\boldsymbol{W}$. This can be decomposed into:

$$P(\boldsymbol{W}) = \prod_{i=1}^{n} P(w_i | (w_1, w_2, \ldots, w_{i-1})) \quad (7)$$

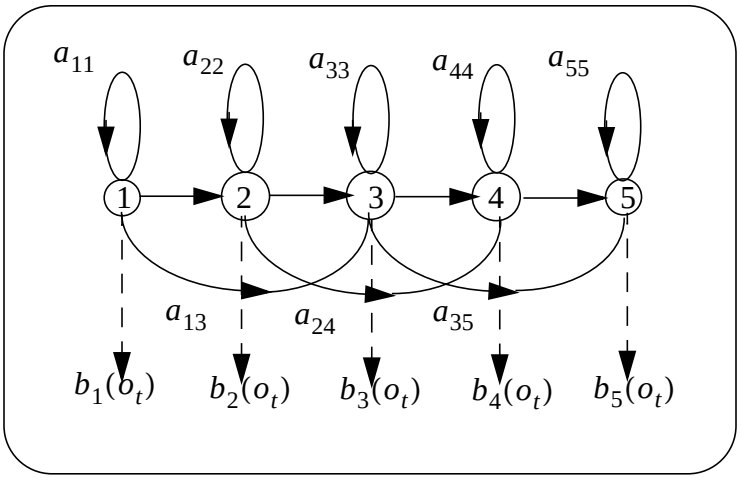


Figure 5. A simple continuous density HMM structure

To estimate the right hand side of the above equation for a moderate vocabulary is itself a very computationally intensive task. Hence, the concept of equivalence classes is used to simplify the above formula. The new equation becomes:

$$P(\boldsymbol{W}) = \prod_{i=1}^{n} Pw(\Phi(w_1, w_2, \ldots, w_{i-1})) \tag{8}$$

where the equivalence class $\Phi(w_1, w_2, \ldots, w_{i-1})$ is determined based on the linguistic knowledge of the language that is being recognized. This is precisely what constitutes language modeling and this is a very task-specific field. A detailed description on building language models for LVCSR systems is given in [15].

Hypothesis search

The final step is to find the desired word string $\hat{\boldsymbol{W}}$ of the acoustic data $\boldsymbol{A}$ using the Baye's formulation. Although, this looks to be a straightforward step, this is the most complex stage of LVCSR systems. The reason for this is the magnitude of the search space. Any algorithm that attempts to do the search process must be extremely optimized in terms of memory and runtime requirements. An excellent review of the search strategies employed in LVCSR systems is given at [16].

## Technical Motivation

Recognition of discrete speech is a more simplified process because of the presence of distinct pauses between words spoken by the speaker. LVCSR systems need to be able to recognize speech that was produced without any prior intention by the speaker

to be transcribed. The should be robust to effects such as background noise, noisy telephone channels, non grammatical speech, emotional speech, etc.

The basic objective of LVCSR systems is the transcription of conversational speech into text, i.e., word strings. Present state-of-the-art LVCSR use complex statistical modeling and pattern recognition techniques to achieve this. The word strings are further broken down into more fundamental speech units called phones and statistical models like Hidden Markov Models (HMMs) are used to represent these phones [17]. These HMM models then need to be trained on large amounts of speech data before they can be used for actual recognition tasks. The purpose of this training is to let the system learn the discriminatory information among the phones.

It should be noted the process mentioned above can also be applied for word models in theory. However in the case of LVCSR systems, there are just too many words to be trained this way. In the case of phone models, only 40 or so models need to be trained. All the words in the English vocabulary can be constructed from these phones. Alternative LVCSR systems have been also built using syllable models (an alternative sub-word speech unit).

Training LVCSR systems is a very important stage of speech recognition. Very often, the performance of a recognizer is hampered by lack of appropriate training data [18]. SWB was intended to be significant constituent of the training data for all state-of-the-art recognition systems. Hence it was critical that the entire SWB Corpus have accurate transcriptions.

Since most of the recognition systems are built using phone models, it would seem that having phone transcriptions for the entire corpus would be the right thing to do. But having humans produce phone transcriptions for such an enormous amount of data is very time consuming and prone to human error. In the early 1980's, a database called TIMIT [19] was transcribed phonetically and was used for smaller recognition tasks. It consisted of 6300 sentences and covered 630 speakers. This was a much smaller database compared to SWB. Various research groups had used this database for evaluating their systems. It was shown that better performance could be obtained by training on the word level transcriptions and allowing the system to learn the phone boundaries. The technique employed was a supervised automatic time alignment approach for finding the phone boundaries [20].

It was decided that a similar approach would be followed for SWB. Having word level transcriptions for the Corpus would be practical and also reduce the scope for human error. The phone level transcriptions could then be produced in an automatic fashion.

The first SWB release was made in 1993 and was welcomed by the speech research community. The next section gives a description of the first release and discusses its strengths as well as drawbacks.

**Summary of the first SWB release**

In 1993, the first SWB release was made by the Linguistic Data Consortium (LDC) [21]. Apart from the audio files, the transcriptions were also provided. Court reporters produced most of the verbatim transcripts, following a manual prepared specifically for the project. It was decided that conversations would be broken at turn boundaries (points

at which the active speaker changed). A flat ASCII representation was used for representing the transcriptions.Their work was checked for formatting errors by an awk script, then twice more by humans during quality control (QC) inspections. These inspections involved checking of spelling errors and speaker identity verification. Each transcript was also accompanied by a time alignment file, which estimated the beginning time and duration of each word in the transcript in centiseconds. The time alignment was accomplished with supervised phone-based speech recognition. The Corpus was therefore capable of supporting not only purely text-independent approaches to speaker verification, but also those which make use of any degree of knowledge of the text, including phonetics. It also could facilitate studies of the phonetic characteristics of spontaneous speech on a scale not previously possible.

## Experimental Conditions

SWB was the first database collected of its type: two-way conversations collected digitally from the telephone network using a T1 line. Use of automatic switching software made it possible to collect the digital version of the speech signals directly from the telephone network, and also to isolate the two sides of the conversations.

SWB was collected without human intervention, under computer control [21]. Interaction with the system was via touchtones and recorded instructions, but the two talkers, once connected, could “warm up” before recording began. From a human factors perspective, automation guards against the intrusion of experimenter bias, and guarantees a degree of uniformity throughout the long period of data collection. The protocols were

further intended to elicit natural and spontaneous speech by the participants. The transcribers' ratings indicate that they perceived the conversations as highly natural.

SWB, in its entirety, consists of 2438 conversations totaling over 240 hours of two-channel data from 541 unique speakers. The average duration of a conversation is six minutes. Of the 500 speakers present in the Corpus, 50 speakers contributed at least one hour of data to the Corpus. Thus, SWB has both depth and breadth of coverage for studying speaker characteristics. Hence, this database was thought to be sufficient enough for extensive training of speech recognition systems.

## Speaker Statistics

The participants' demographics, as well as the dates, times, and other pertinent information about each phone call, are recorded in relational database tables. A lot of criteria were taken into consideration while recruiting the speakers for the data collection task. It was intended that the talkers be broadly representative of adult speakers of American English. Serious effort was undertaken to cover various dialects too. The speaker distribution based on dialect classification is given in Table 1. The demographic coverage of the dialects is shown in Figure 6. A complete description of the speaker selection procedure and the calling protocol is given in [21].

## CD Distribution by LDC

The complete release consisted of 25 CD-ROM's. Each conversation had a four digit identification number. The complete set of transcriptions were on one CD-ROM. The orthographic transcription files were named as swXXXX.txt and the time aligned

Table 1. Number of callers per dialect area

| Dialect Area | Count |
|---|---|
| South Midland | 155 |
| Western | 85 |
| North Midland | 77 |
| Northern | 75 |
| Southern | 56 |
| New York City | 33 |
| Mixed | 26 |
| New England | 21 |

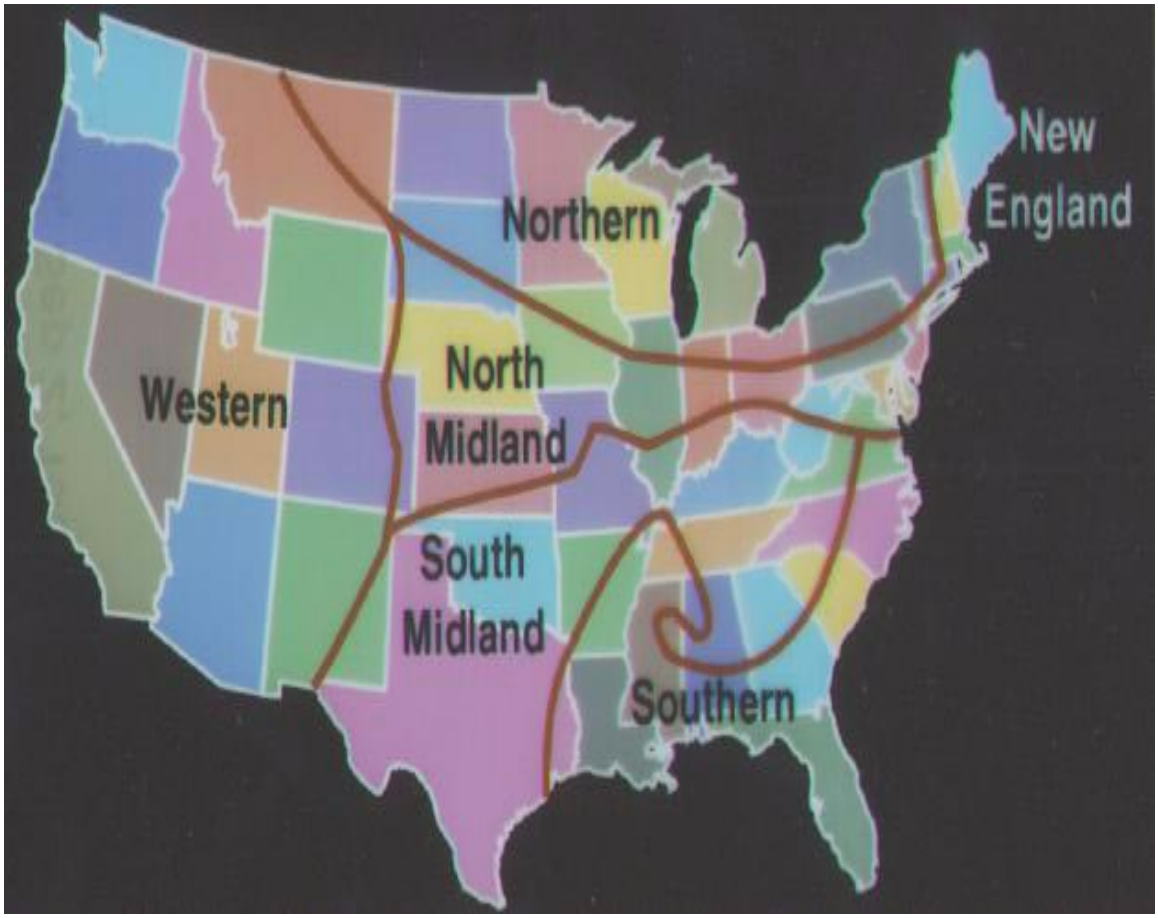


Figure 6. SWB dialect map

transcripts were named as swXXXX.mrk, where XXXX corresponded to the conversation number.

The audio files were distributed in the NIST wav format. Each audio file had a standard 1KByte header. An example header is shown in Figure 7. The transcription files also had some header information that had information regarding the topic ID, transcriber ID and the other related information. The first few lines of a typical transcription file are shown in Figure 7. The time alignment files were arranged in fixed records of four fields, where the first field was the speaker (A or B), the second was the estimated start time in seconds, the third was the estimated duration in seconds, and the fourth was the word whose start time and duration are estimated. The first few lines of a typical mrk file are shown in Figure 9.

```
speaker_id1 1423
speaker_id2 1662
recording_date 920508
recording_time 2204
conversation_id 4940
database_id SWB1
channel_count 2
sample_max1 4015.500000
sample_max2 4015.500000
sample_coding mu-law
channels_interleaved TRUE
sample_count 4798496
sample_rate 8000
sample_n_bytes 1
sample_sig_bits 8
```

Figure 7. Example of a NIST wav file header

```
FILENAME:       4940_1423_1662
TOPIC#:         302
DATE:           920508
TRANSCRIBER:    nk
DIFFICULTY:     2
TOPICALITY:     1
NATURALNESS:    2
ECHO_FROM_B:    1
ECHO_FROM_A:    1
STATIC_ON_A:    2
STATIC_ON_B:    1
BACKGROUND_A:   2
BACKGROUND_B:   2
REMARKS:         None

================================================

A: Okay [children].

B: Okay Carol. So, air quality.

A: Yeah. Is it, [noise] {sounds like water running
and she is doing dishes} I know in here, uh, down-
town Dallas, it's, you, I mean you drive by and you
can just, you can see it.

B: Uh-huh.
A: But, then again [throat_clearing] I originally
was from California and, uh, there is a big differ-
ence between Texas and California. #And, uh# --

B: #Surely.#

A: -- they'd have their smog alerts and where you'd
have to stay indoors for so many hours with an air
conditioner. And, of course, they don't have that
```

Figure 8. An example transcription file

```
A       0.04    0.42 Okay
A          *       * [children].
B       0.82    0.22 Okay
B       1.06    0.34 Carol.
B       3.58    0.34 So,
B       3.92    0.20 air
B       4.12    0.70 quality.
A       5.40    0.22 Yeah.
A       6.16    0.16 Is
A       6.32    0.16 it,
A          *       * [noise]
A          *       * {sounds
A          *       * like
A          *       * water
A          *       * running
A          *       * and
A          *       * she
A          *       * is
A          *       * doing
A          *       * dishes}
A       7.02    0.10 I
A       7.12    0.22 know
A       7.34    0.08 in
A       7.42    0.30 here,
A       7.80    0.22 uh,
A       8.36    0.44 downtown
A       8.80    0.46 Dallas,
A       9.26    0.22 it's,
A       9.60    0.20 you,
A       9.82    0.10 I
```

Figure 9. An example time alignment file

Apart from the above files, some anciliary text files were also included in the release. These were mainly a compilation of tables that included various statistics about the speakers, topics etc. Care was taken to ensure that the privacy of the callers and their phone numbers is maintained. This information was not included in the release. A

dictionary was also developed as a by-product of the automatic time alignment procedure but was not included in the first SWB release. Pronunciations for all words were also included in this dictionary.

## Changes to the original distribution

A lot of care was taken while collecting the SWB database. A great number of technical issues related to the hardware implementation as well as the software requirements of the collection protocol were addressed. But, in spite of all precautions and guarantees, a few technical problems did occur with the collection system. These could be classified into two major categories:

- Digital Noise (also called “static”)
- Loss of synchrony between the two telephone channels

The first problem, namely static noise was caused due to hardware failure. One of the four telephone interface cards used began to fail intermittently. Data being collected from this channel was being replaced by random values. This caused the speech to be heard as very loud static. This happened for periods ranging from a few samples up to several seconds. An example speech file containing static is shown in Figure 10. The occurrence of the aforementioned problem can be seen in the spike that occurs in the

Figure 10. Occurence of “static” in speech

center of the figure. A retroactive solution to this problem was to mark the transcriptions with a “[static]” tag whenever it occurred. Conversations for which this problem was significant were dropped from the database.

The second category deals with the loss of synchrony between the two channels of the conversation being recorded. These problems were very subtle and could not be detected early enough in the collection process. These could further be divided into three categories. All categories had a common symptom in the fact that the there existed a time lag between the speech signal on one side of the conversation and it’s echo on the other side. A description of these three categories follows

- The first problem was an asynchronous startup of recording between the two channels. In the original specifications, “simultaneous” startup was called for, but the need for scientific precision was not understood by the applications programmers. As a result the recording of one side of each call was being started either 55, 110 or 165 ms after the other side.

- The second problem was a more serious one but occurred rarely. The computer being used for recording conversations was sometimes overloaded and was unable to record parts of the conversation. This manifested itself in loss of speech data on some channels. Three conversations were found to have this problem and were removed from the database.

- The third problem was small changes in synchrony between type two channels, due to a pseudorandom dropping of 2 ms chunks of data on either side. Over the course of a 10 minute conversation, these could accumulate to a differential of 30 or 40 msec between sides--enough to change a cross-channel echo from inaudible to audible, for example, or from barely audible to very noticeable, for a human listener. The dropping of the data was caused due to faulty code written for the collection protocol.

Corrections for these problems were done at the National Institute of Standards and Technology [22]. Both sides of the problematic conversations were compared with a cross correlation measure at various delays. The lag time which showed the best peak in

the correlation function between speech on one side and its echo on the other was measured throughout the file. If this occurred in the initial portion of the speech file, the problem was perceive to be of the first category. If this occurred in the later parts of the speech file, it was considered evidence of the loss of data in 2 ms chunks. The rectification process included addition of appropriate silences.

### Echo cancellation

Initial attempts to transcribe SWB had not dealt effectively with the echo present in the audio data. This had caused numerous problems with swapped channels in transcriptions and with incorrect transcriptions (because the amplitude of the echoed speech is often on par with that of the speech data from that channel). To avoid these problems and to provide the validators with the highest possible audio quality, all conversations needed to be echo cancelled before transcription. This process consists of simply passing the data through a standard least mean-square error echo canceller [23]. Figure 11 shows an example. The top figure is an example of a audio file that has a lot echo. The resulting audio sample after echo cancellation is shown the bottom figure. It is obvious that echo cancellation does help in making the job of the validator easy. Echo cancellation was appropriate for most of the audio files in the SWB Corpus. But it had a reverse effect for a few audio files and it was decided that the non echo canceled versions of the audio files would be used for these.

## Transcription and Segmentation problems

The original LDC transcriptions were segmented by conversation turn boundaries. The transcriptions themselves were word-level transcriptions as described in the earlier

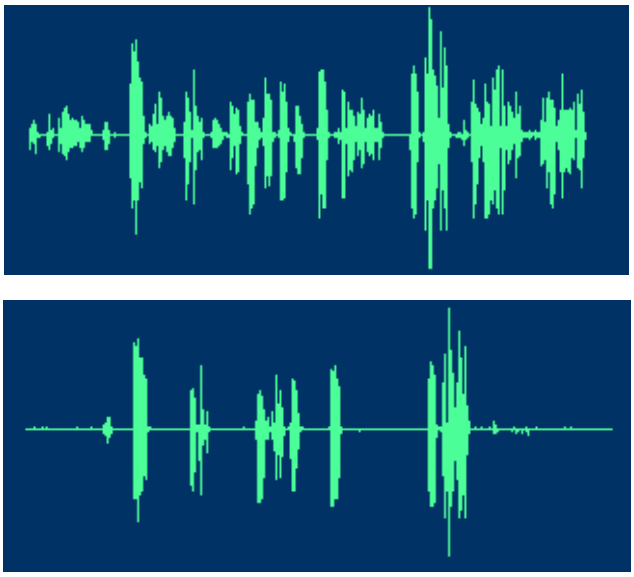

Figure 11. An example of a speech file before and after echo cancellation

section. For a task as huge as SWB, the conventions associated with transcriptions are always highly controversial and application specific. SWB was widely being used for a variety of speech research tasks. The original transcriptions were proving to be unsuitable for many such tasks. For example, the original transcriptions were not case sensitive and this made the database unsuitable for natural language processing experiments [24]. Another example was the lack of classification of disfluencies (words such as “uh” “hm” uh-huh”) in the original transcriptions. This was a hindrance to research tasks that included disfluency analyses [25].

Many research groups have tried to rectify minor issues with the transcriptions [26]. The final outcome was that numerous versions of the SWB Corpus were floating around; few of these improved transcriptions were folded back into the original LDC distribution but a lot of research dollars had been spent in vain trying to clean up the transcriptions.

In 1998, the Institute for Signal and Information Processing (ISIP) at Mississippi State University began a comprehensive project to substantially improve the quality of the segmentations and transcriptions. The main goal was to incorporate many of the incremental changes made in previous work and to provide a much cleaner database. The next few chapters describe the workflow process adopted and the final outcome of this project.

# CHAPTER III

# SEGMENTATIONS AND TRANSCRIPTIONS

Performance improvements of LVCSR systems have now become less dramatic. The technology has attained a stage, where even incremental improvements in performance are considered a success. Many research groups are focussing on issues related to the quality of the database. Unfortunately, the SWB corpus had a lot of problems related to the segmentations and transcriptions and casual reviews processed by many sites revealed that most of the transcriptions are unsuitable for training LVCSR systems.

As stated earlier, ISIP began a project to clean up the SWB segmentations and transcriptions in 1998. The motivation for this project dates back to a pilot experiment conducted during the 1997 Summer Workshop (WS'97) at the Center for Language and Speech Processing (CLSP), Johns Hopkins University. This study showed that improved segmentations and transcriptions resulted in improved acoustic models. Simply retranscribing the test database resulted in 2% reduction in the Word Error Rate (WER) [27].

This chapter will discuss the original SWB segmentation and transcription guidelines and suggest and improved set of guidelines that were used by ISIP. The workflow process for the project as well as the software developed by ISIP will be summarized. A variety of problems that highlight the problems associated with

conversational speech and were observed during the course of this project will also be discussed. These have been nicely summarized in a FAQ [28].

## Original Guidelines

This section provides a description of the original guidelines used while transcribing the SWB Corpus. It first provides a description of the approaches used for segmenting and transcribing speech databases and the motivation for the original guidelines.

### Segmentations

Historically, speech segmentation has been guided by either linguistic or acoustic metrics independent of the other. In linguistic segmentation of data, boundaries are placed at the end of meaningful phrases or conversation turns whereas in acoustic segmentation, they are placed in acoustic silence between words. Both these metrics have some advantages as well as drawbacks.

Linguistic segmentation results in utterances that contain meaningful phrases and hence can be used in the training of a language model. This approach is preferred for tasks which need a robust LM. It is effective in maintaining clear linguistic context, but suffers from two major drawbacks. First, placing the boundary based on linguistic meaning results in some boundaries between words with very little silence in between. This results in cutting off of the beginnings and ends of the words and adversely affects the training of acoustic models. The second drawback relates to the length of the resulting utterances. During conversational speech, the speaker often speaks for 15-20 seconds elaborating on the same phrase and with no pauses. Thus, linguistic segmentations sometimes results in

utterances that are much longer than the ideal 10 second utterances. An example SWB utterance that is very long is shown in Figure 12. Training models on very long utterances is not effective. In fact, the utterances used for evaluations are much shorter than 10 seconds and hence long utterances are not preferable.

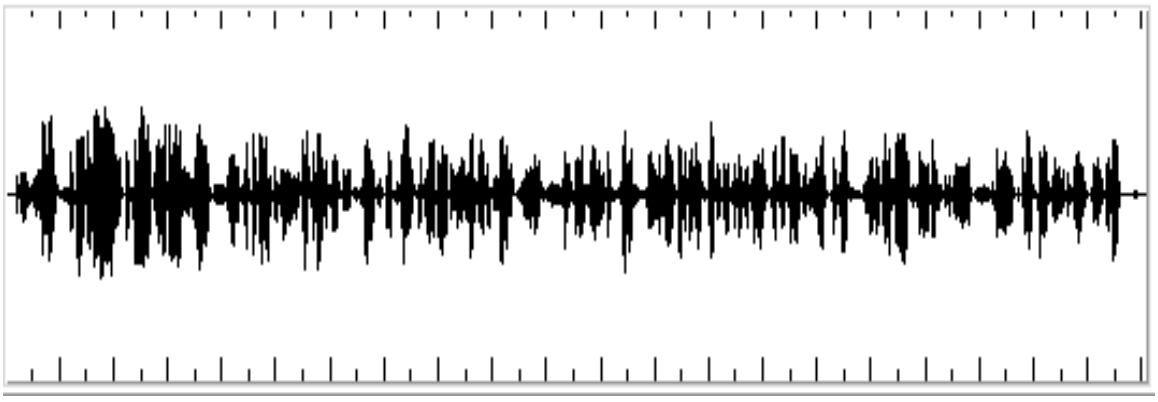

Figure 12. In the above waveform, a speaker provides 21 seconds of continuous speech

Acoustic segmentation of speech has it's advantages too. Placing boundaries in regions of silence (or very less energy) is necessary for training the acoustic models effectively. But the major drawback here is that the utterances lack linguistic context and hence are of not much use in training the language model. As stated earlier, the use of a proper LM also greatly enhances performance and the lack of it hinders improvement in performance.

The original segmentations for the SWB corpus were done at conversation turns. This was equivalent to linguistic segmentations and resulted in utterances with good linguistic context but poor acoustics at the beginnings and ends. The utterances consisting

solely of silence were discarded. Also, due to the nature of the conversations, most of the segmentations occurred between two words with not much silence in between. Also there were no guidelines for separation of speech from non-speech events such as laughter and noise. Very often, noise or laughter was included along with speech in the same utterance when they could have been separated. This also affects the training of acoustic models.

## Transcriptions

The original LDC transcriptions were word-level orthographic transcriptions. The phone level transcriptions were later obtained in an automatic manner. Around half of the conversations in the SWB corpus were transcribed by court reporters and other half was transcribed by validators employed by TI. Verbatim transcription of the conversations was preferred. A detailed description of the original guidelines can be found at [29]. A few of the interesting issues are described below:

- Verbatim transcription was followed. This meant that none of the grammatical errors in the transcriptions were corrected.
- Pronunciations: A dictionary form was used and imitation of pronunciations was not allowed. Hence the corresponding phone models were not very accurate for words where the speaker uses alternate pronunciations.
- Word abbreviations were avoided. For example, Fort Worth was preferred over Ft. Worth.
- Some punctuations were allowed. All sentences ended with a period and commas were used to note changes in grammatical structure of the speech.
- Contractions were discouraged. This again meant that the phone models were built on wrong pronunciations.
- The first release did not provide a corresponding lexicon.

| | | | |
|---|---|---|---|
| [TV] | [chiming] | [music] | |
| [baby] | [clanging] | [noise] | [squeak] |
| [baby_crying] | [clanking] | [nose_blowing] | [static] |
| [baby_talking] | [click] | [phone_ringing] | [swallowing] |
| [barking] | [clicking] | [popping] | [talking] |
| [beep] | [clink] | [pounding] | [tapping] |
| [bell] | [clinking] | [printer] | [throat_clearing] |
| [bird_squawk] | [cough] | [rattling] | [thumping] |
| [breathing] | [dishes] | [ringing] | [tone] |
| [buzz] | [door] | [rustling] | [tones] |
| [buzzer] | [footsteps] | [scratching] | [trill] |
| [child] | [gasp] | [screeching] | [tsk] |
| [child_crying] | [groan] | [sigh] | [typewriter] |
| [child_laughing] | [hiss] | [singing] | [ugh] |
| [child_talking] | [horn] | [siren] | [wheezing] |
| [child_whining] | [hum] | [smack] | [whispering] |
| [child_yelling] | [inhaling] | [sneezing] | [whistling] |
| [children] | [laughter] | [sniffing] | [yawning] |
| [children_talking] | [meow] | [snorting] | [yelling] |
| [children_yelling] | [motorcycle] | [squawking] | |

Figure 13. A list of typical non-speech sounds that were used in the original transcriptions

- Various tags for non-speech sounds were used. A list of these tags is given in Figure 13.

- No conventions were used for partial words. The occurence of partial words is very high in conversational speech and hence would have been beneficial to have these.

## Changes in guidelines

In an effort to clean up the original SWB segmentations and transcriptions, ISIP has developed a revised set of guidelines. This section will describe the motivation for these revisions and their advantages. The first part of this section deals with the segmentations and the latter discusses the revisions made to the transcriptions.

## Resegmentation

As described in the earlier section, there are two primary ways to segment a speech database: linguistic segmentation vs. acoustic segmentation. Both of these approaches have advantages as well as drawbacks. Hence, the best approach would be one that strikes a balance between these competing paradigms: manually placing boundaries where there is acoustic silence, maintaining linguistic context, and regulating the length of the utterances. A similar approach for segmentations on the test set gave a 2% absolute improvement in WER [27].

Resegmentation is a 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. The strategy used in this work is as follows:

- Segment at locations where there is clear silence separating each segment
- 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. These general guidelines were modified to produce the short set of specific guidelines shown below [30]:

- Each utterance should be padded by a nominal 0.5 second buffer of silence on both sides. In general, these silence buffers can range from 0.35 to 0.75 seconds. This provides ample silence at the start and end of utterances to negate the possibility of acoustic information being truncated.

- The boundary can only be placed in a "silence" consisting solely of channel noise and background noise. Whenever possible place the boundary in a section with very low energy (visually speaking, this is a flat part of the signal.) It is the intention of this work to have "clean" utterances where each boundary is in a point of silence, each utterance is buffered by silence, and each utterance contains a meaningful phrase. Boundaries in noise locations cause corruption of delta features leading to less accurate acoustic models.

- The 0.5 second buffers can contain breath noises, lip smacks, channel pops, and any other non-speech phenomena. However the boundary can not be placed in a noise of this sort.

- No utterance can be longer than 15 seconds. As an utterance approaches 15 seconds in length, the validator is allowed to find a point of segmentation that will generate silence buffers less than 0.5 seconds but not less than 0.1 seconds. This rule ensures that the data generated is suitable for use in speech recognition systems. Utterances longer than this can produce a search space which extends beyond the capability of common computers to deal with efficiently.

- Every utterance containing only silence must be greater than 1.0 seconds in duration. Otherwise the silence region could be used as part of the buffer for the previous and next utterances.

## Retranscription

The retranscription phase of the SWB corpus was mainly concerned with correcting the original LDC transcriptions and providing guidelines for handling laughter word and partial words. A complete description of our modified transcription conventions is given at [30]. Many of these rules were a by-product of problems pointed out by the validators. Each time that a validator was not able to easily arrive at a

transcription by following the conventions, new rules were added a rule to help maintain clarity and consistency. A detailed description of these issues is given in [30].

## Workflow process

The workflow process developed at ISIP for cleaning up the SWB Corpus was an incremental process where the segmentations were first completed and the transcriptions were done later. Initially, one validator would both segment and transcribe the data, and, when a large portion of data was ready for release, the project manager would run a small number of quality control scripts to verify the validators work. This approach was flawed in a couple of respects. First, the data was only reviewed in large chunks (on the order of 100 conversations) which meant that the same type of error may have been propagated through a large number of conversations before being corrected. Lastly, the validators had difficulty focusing on both the segmentation and transcription because the number of issues involved in each is substantial.

The improved workflow process schematic is given in Figure 14 [31]. The first benefit this new process provides is an increased review of the data. Each conversation is now completely reviewed by two different validators. One validator only resegments the data — building a set of utterances which match the specifications of our transcription guidelines. The other validator concentrates on making transcription corrections and makes note of any segmentations that are questionable so they can later be reviewed by the project manager. This segmental approach has worked well because the validators are able to focus on a single task rather than balancing both segmentation and transcription.

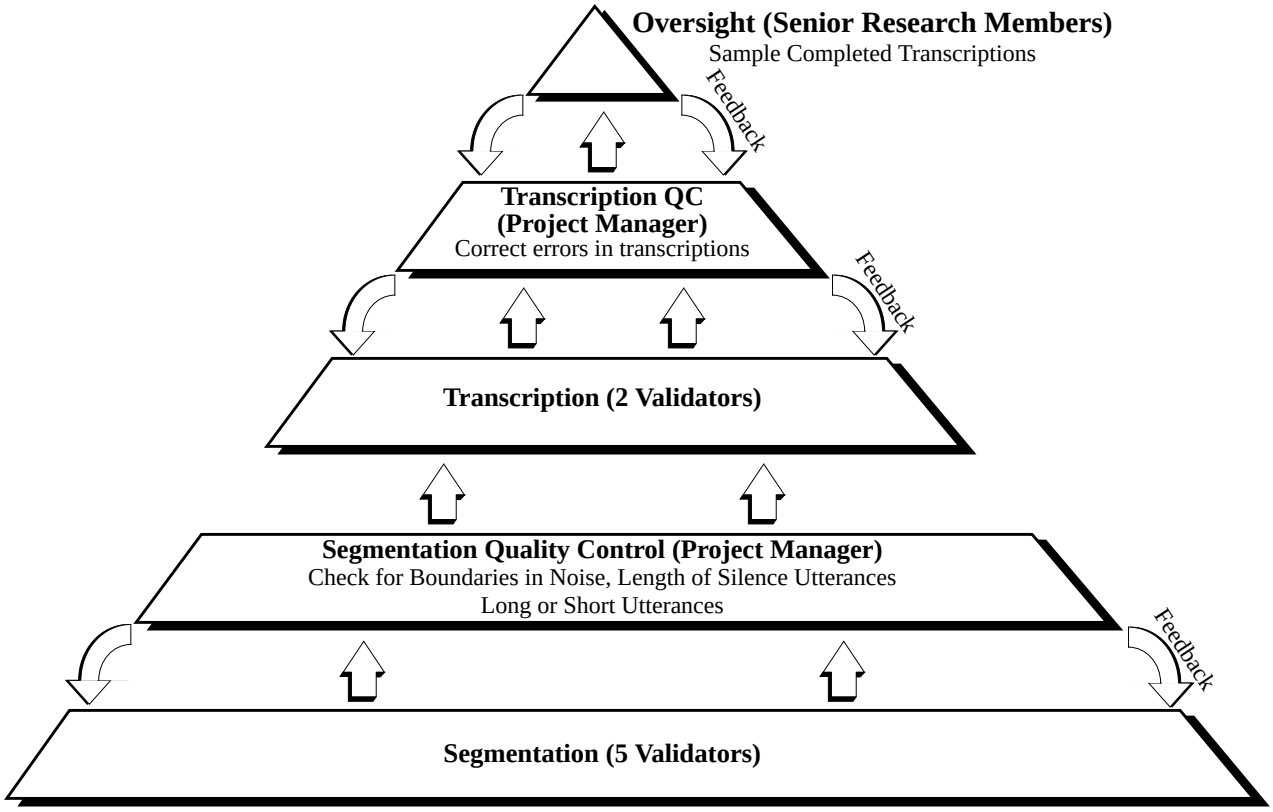


Figure 14. New workflow for segmentation and transcription

## Segmenter tool

A segmentation tool [31] was developed by senior Ph.D. students at ISIP as part of this work. It is a point-and-click interface tool designed to streamline 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. The underlying principle is that all speech data must be accounted for. This helps in explicit marking of silences and no audio data is missed by the validator. A screenshot of this tool is given in Figure 15.

The tool also helps reduce the training period for the validators. The interface helps them to get accustomed to the segmentation/transcription process very quickly.

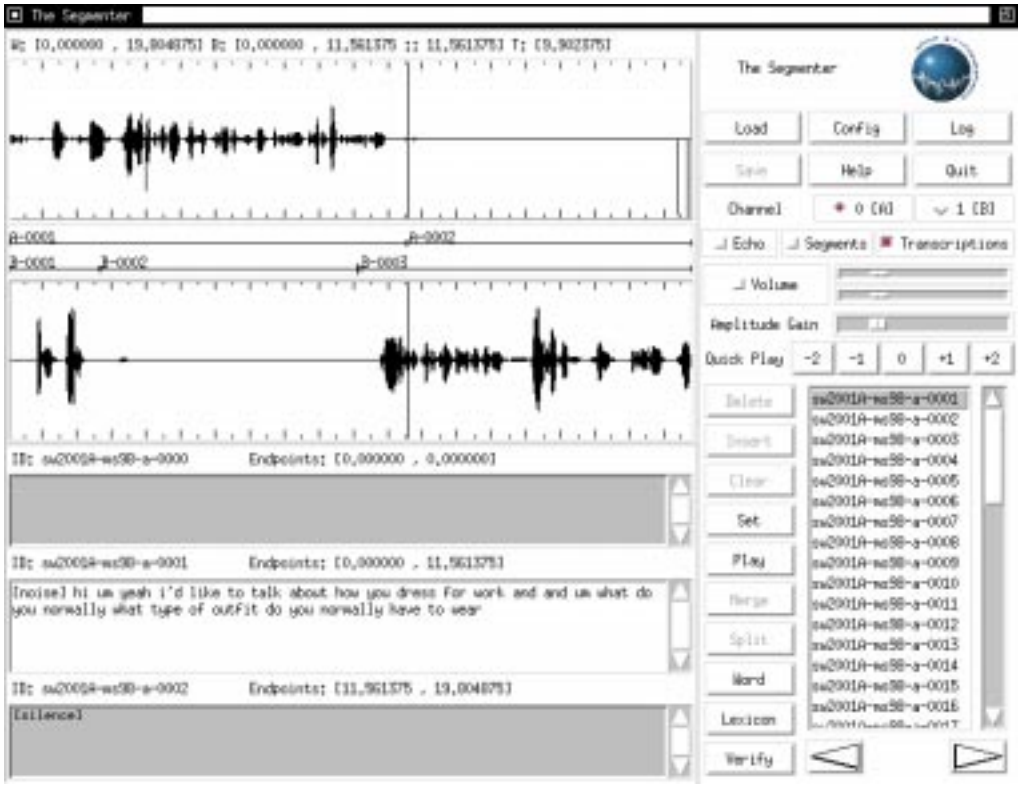


Figure 15. Screenshot of the segmentation tool.

Changing of boundaries was as easy as moving the mouse to the right position. Transcription changes could also be done very efficiently. The ability to listen to two channel data helped differentiate echo from the actual signal. Properties such as zooming in and.or playing selected regions of speech were also incorporated. In summary, this tool was very helpful for the timely completion of this project.

## SWB FAQ

This section will give a detailed description of the important problems associated with transcribing conversational speech and provide guidelines for tackling them. These were encountered during the retranscription phase of the project. The FAQ page

maintained by ISIP has the complete list of such problems as well as the consensus reached to transcribe them [28].

Many of the issues described in this section highlight the complexity of conversational speech recognition. 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 the validator must make what amounts to an educated guess. These problems result in most of the final word errors in the revised data. 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 was the practice in this work to transcribe all speech in the database with the most likely word given all of the available information.

- Title capitalizations: We used standard grammar rules and capitalized the first word, last word, and kept prepositions under five letters lower case (example: “Gone with the Wind”).

- Compound words: 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: Massachusetts should be called Taxachusetts). The convention on these words, called coinages, was to transcribe the word in braces — in this case, “{Taxachusetts}”.

- Mispronunciations: Occasionally speakers mispronounce a word or say a word they didn’t mean and then correct themselves (example: I blame the splace space program). The convention decided was to 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: There are 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. In such cases, the convention is transcribe then 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-”). The 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 a rule was made 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]”.

- Asides: A situation that occurs relatively infrequently in SWB is when one of the two speakers in the conversation talks to a person in the background. In the past, this may have been transcribed as [noise], as part of the normal transcription, or, worse, not transcribed at all. This could have dire consequences for training or testing a system since the acoustics for these “asides” would be on the same level with the conversational acoustics. Also, these asides will often carry over into the conversation between the two primary speakers. The practice was adopted of transcribing the parts of the conversation spoken as asides between the markups “<b_aside>” and “<e_aside>”. (example: “excuse me <b_aside> i said go outside and play <e_aside> sorry”)

# CHAPTER IV

# HUMAN TRANSCRIPTION PERFORMANCE

It is a widely known fact that humans outperform machines in speech recognition tasks by orders of magnitude [32]. This has been bench marked on some conversational speech databases [33]. Unlike machines, humans perform well on this task irrespective of the domain or Signal to Noise ratio (SNR) settings [33]. Understanding how humans recognize speech will help improve the performance of machines to a large extent.

Transcription of speech databases is largely done by humans due to need for accurate transcriptions. For the SWB corpus, it was mentioned earlier that word level transcriptions are the better option. A detailed review of a small section of the original transcriptions revealed that, on average, 8% of the words transcribed are in error [34]. Thus it was necessary to retranscribe the SWB Corpus. Asserting the quality of these transcription is equivalent to assessing the performance of humans on recognizing continuous speech.

In this chapter, we examine the WERs of the new SWB transcriptions, and describe a quality control (QC) procedure to improve their accuracy. The WERs before and after the quality control process are analyzed. Issues pertaining to human recognition performance are then discussed and suggestions to improve machine performance are given.

Table 2. Comparison of error rates for the LDC and revised transcriptions

| Transcriber | WER |
|---|---|
| LDC | 5.4% |
| ISIP before QC | 3.7% |
| ISIP after QC | 1.5% |

## Preliminary Experiments

The results of an experiment to ascertain the improvement in quality of the new transcriptions is detailed in Table 2 [34]. The conversation used was sw2137. It was a 4 minute conversation and had a difficulty level of 2 on a scale of 5 (5 being hardest).0V validators transcribed data from the same segmentation of conversation and were scored against a reference that was also transcribed from that segmentation. The errors shown in the table are significant errors which only include deletion, insertion, or substitution of a word. These specifically do not include minor differences in partial words, differences in transcription conventions (when scoring the LDC data), and marking of noises. One can see from the table that the revised transcriptions better the LDC transcriptions by a significant margin. Also, the reduction in WER after the QC process is significant. This highlights the need for a stringent QC process. The next section describes the evolution of the QC process at ISIP and the subsequent sections analyze the quality of the final transcriptions.

**The Quality Control process flow**

Resegmentation and retranscription of the SWB corpus involved manual validations. As with any process involving manual validation, this process is subject to occasional error. To combat this problem, a stringent QC process was developed.The process evolved based on the feedback received from the validators and review of transcriptions during the initial phase of the project. An overview of this process is shown

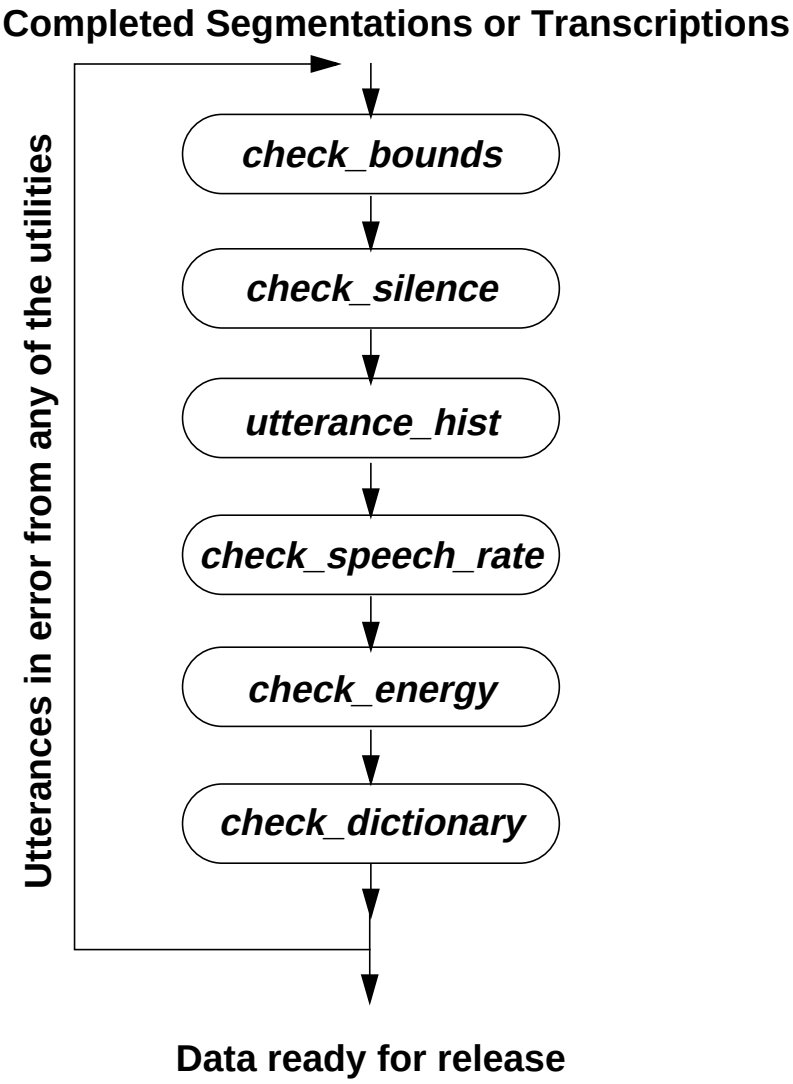


Figure 16. Sequence of quality control utilities used to check for segmentation and transcription errors

in Figure 5. The process was run for every single conversation in the corpus before it termed as finished. Batches of 20-40 conversations were QC'ed on a weekly basis.

In this, problem utterances are marked by the quality control scripts and each of these is reviewed and corrected if necessary. At this point any questions logged by the validators regarding segmentation and transcription are reviewed and decided upon via interaction with the research community. Beyond this set of internal reviews, an incremental release of the data was made to the public domain for review by speech technology experts who played a part in the quality control process via their feedback.

At the core of the quality control regimen are a set of utilities that automatically tag utterances that have common errors such as misspellings and boundaries in noise. The sequence of scripts used is shown in Figure 6. Notice that the process is iterative as each marked problem must be adjudicated before the conversation is released. Each of the utilities are described below:

- check_bounds: In the early stages of the project the validators were not protected by the segmentation tool from mistakes such as putting the right boundary before the left boundary. The check_bounds utility will find all such gross errors in the boundary alignment. The utility verifies that every sample of data in the speech file is accounted for by the transcription start and end times. It does so by making sure that the start time of every utterance (or word in the case of word alignment files) is equal to the end time of the previous utterance or word. It also checks that the end time of the last utterance or word is equal to the last sample in the file and that the start time of the first utterance is zero.

- check_silence: One of the new transcription conventions was that every utterance marked as containing only silence should be at least 1.0 second long. At times the validators intended to merge a pair of utterances but unintentionally left a dangling silence-only utterance which is extremely small. This utility finds these problems by tagging all utterances that are transcribed as "[silence]" but are shorter than a specified minimum duration. For our quality control process, the minimum duration was set to 1.0 seconds. In some

circumstances, silences less than 1.0 second are allowed, but all such utterances would have been reviewed.

- utterance_hist: From experience, the belief was that the average SWB utterance should be between 6 and 8 seconds long and should rarely be greater than 15 seconds or less than 2 seconds. In reviewing data, it was found that the validators were not paying close attention to these parameters. This utterance_hist utility accepts a list of transcription files and for those files flags those utterances whose duration falls outside of the accepted range (2 secs - 15 secs). It also produces comprehensive statistics for that list of files including: number of conversations processed, number of non-silence and silence-only utterances, number of words, hours of non-silence and silence-only data in the conversations, mean duration of non-silence utterances, standard deviation of duration among non-silence utterances, and maximum and minimum utterance lengths. These statistics are used to characterize the data being produced and to search for any trends in the data which would lead to identify problems in the new transcriptions.

- check_speech_rate: It was found that most gross errors in transcriptions such as accidentally replicating part of the transcription twice in one utterance can be easily found by examining the speech rate of each utterance. This is a measure of the number of words transcribed per second of speech in the utterance. It was also found that a vast majority of correct utterances have rates between 0.5 and 5.0 words per second. Thus, this quality control script flags any utterances which have speech rates outside of this range. There are, of course, utterances which are in error yet still fall within the range of accepted rates. The number of these was minimal in the released data and could be corrected in the later stages of the project.

- check_energy: This utility was the primary means for verifying that the validators are following the rules for placing boundaries in a low-energy area. The utility uses a standard algorithm [35] to determine the nominal channel energy level. For each utterance in a conversation, check_energy finds the average energy of a window around the boundary. If that average energy is larger than the noise floor of the conversation by a certain amount (typically 25 dB) then the boundary is flagged as occurring in an impulsive noise. This method has been extremely successful in finding boundaries placed in noise or echo and has helped in demonstrating to the validators examples of correct and incorrect boundary placement.

- check_dictionary: A revised dictionary was built from the improved transcriptions [31]. This dictionary provides a pronunciation for each word in the conversations. With each corrected transcription comes words that are

currently not in the dictionary — these are usually partial words, proper names, or laughter words. check_dictionary is used to find those words that are not in the dictionary. Each of these are individually reviewed and, if the word is correct in the transcription, are added to the dictionary. This helps in finding any misspelled words or misused words. Using this utility is not foolproof since words can be mistranscribed in the transcription though they do appear in the dictionary. An example of this is a transcription of "World War I" which should be transcribed as "World War One", but since "I" is in the dictionary, check_dictionary will allow this phrase to pass.

- get_val_stats: One of the best indicators of progress in reframing the transcription and segmentation procedures has been the increased performance of validators accompanied by an increase in accuracy. get_val_stats is used to generate statistics on a per-validator basis. With this utility, one can determine the hours of data transcribed, the number of conversations completed and the real-time rates of the validators over a given period of time. It was found that daily feedback to the validators on their real-time rates and data production has been a great motivator for them to continue to work hard.

## Cross-Validation

Cross-validation experiments are very important indicators of the quality of the new transcriptions and segmentations. In these tests, a number of validators segment/ transcribe the same conversation and their work is compared against a reference segmentation/transcription. The reference is reviewed by a set of experienced speech researchers before it is ready to be termed as correct. Also, this is a blind test and hence the validators do not know that they will be scored on that particular conversation. The conversations used for these tests were carefully chosen to be typical of average Switchboard conversations. Issues such as presence of echo, background noise and channel noise were also considered in choosing the conversations. The next two sub-sections show cross-validation results on segmentations and transcriptions.

## Segmentation cross-validation

Cross-validation experiments for segmentations were performed in order to review the new segmentations for possible errors that could not be rectified by the QC scripts. Most of these errors are related to finding the “more meaningful phrase” among two choices with similar acoustic boundaries. The results from a typical experiment are shown in Table 3. The error rate was calculated as:

$$E = \frac{w}{s} \tag{9}$$

where:

$E$ : the error rate

$w$ : total number of wrong segmentations

$s$ : total number of segmentations for the conversation

The conversation used (sw3093) for the cross-validation experiment was typical of most of the conversations in the SWB database with respect to the channel noise and the

Table 3. A typical segmentation cross-validation experiment

| Segmentation Source | Error Rate |
|---|---|
| WS’97 | 7.5% |
| ISIP: validator A | 1.2% |
| ISIP: validator B | 1.4% |
| ISIP: validator C | 0.4% |

presence of echo. The echo-cancelled version of the audio file was used. The duration of the conversation was 6.8 minutes, which is also close to the overall mean conversation duration.

Any boundary that is in the same area — within 0.2 secs — of the reference location and is not in a high energy region of the signal was considered to be a correct segmentation. The errors observed in this test were mainly due to less than desirable linguistic segmentations. A best split is termed as one that makes sense with respect to the linguistic content as well as the acoustic content of the utterance. The errors observed were basically related to merging or splitting a couple of utterances based on the linguistic content (phrase structure). Proper feedback was given to the validators based on this test.

The comparable WS'97 segmentations were also compared to these same reference transcriptions. The results show that the present ISIP segmentations are more consistent with the conventions and that the data generated by the validators is consistent. The decrease in error rate from over 7% to 1% is very significant, and is one reason for vastly improved automatic word alignments with the new data.

Similar cross-validation tests were performed on a monthly basis during the segmentation phase of the project. The results from these tests were comparable and have been documented in the quarterly project reports [36 37 38].

### Transcription Cross-validation

Transcription cross-validation is much more important to assess the performance of the validators. The retranscription part of the project started in May 1999. Since, then cross-validation tests were performed on a monthly basis. The conversations were chosen

Table 4. Cross-validation results for SWB transcriptions.

| Validator | Word Error Rates (%) | | | Major Error Modality |
|---|---|---|---|---|
| | sw4928 | sw3311 | sw3467 | |
| A aragh | 1.56 | 1.81 | 1.53 | partial words |
| B george | 1.93 | 1.98 | 2.04 | typos and laughter words |
| C maddox | 2.15 | 2.16 | 2.20 | capitalization |
| D vogel | 1.41 | 1.52 | 1.70 | partial words |

based on factors that included speaker dialect and speaker rate in addition to the ones discussed earlier. The results for three such conversations are show in Table 4. The WER was calculated as follows:

$$e = \frac{S + I + D}{W} \tag{10}$$

where:

$e$ : word error rate

$S$ : number of substitutions with respect to the reference transcription

$I$ : number of insertions with respect to the reference transcription

$D$ : number of deletions with respect to the reference transcription

As is evident from the table, error rates were in the range of 1.75%. Very few of these errors were cases where a word clearly spoken in the utterance was misses by the transcriber. Most of the errors involve violation of the conventions for a partial word or laughter word transcription. Once a cross-validation test was complete, validators were given feedback in an effort to improve their performance. However, it does appear that the

experienced validators had approached a human limit on performance. According to the strict error measures used in Table 4, it was difficult for a validator to reduce the error rate below 1.25%.

## Analysis of Cross-validation results

The cross-validation tests provided us with data to examine transcriber error rates and agreements. As sated earlier, human transcription performance is a good indicator of speech recognition performance of humans. Unlike in many other human recognition experiments, the transcribers were not under time pressure and could back up and re-play any portion of the recording and review their transcriptions at any time. They were also at liberty to use dictionaries and other "performance-enhancing" tools. So, theoretically, they should have achieved the best possible performance. This section analyses the transcriber performances and their errors and highlights the subjectivity involved in generation of "clean" SWB transcripts.

### Transcriber agreement

In order to compute transcriber agreement, the SCLITE [39] scoring package was used. The transcribers were said to agree where each of them supplied identical words which aligned to each other. Since raw counts of agreed-upon words would be difficult to interpret, we produced a percentage by dividing the sum of agreed-upon words by the total number of words in the reference transcription. Figure 17 summarizes the inter-transcriber agreement between transcribers for the conversation sw4928. The over-all 4-way agreement (T1 vs T2 vs T3 vs T4) percentage of 98% indicates a very high human recognition rate. A review of the cross-validation transcriptions also shows that the

differences are mainly with regards to partial words. These differences are tabulated for the conversation sw4928 in Table 5.

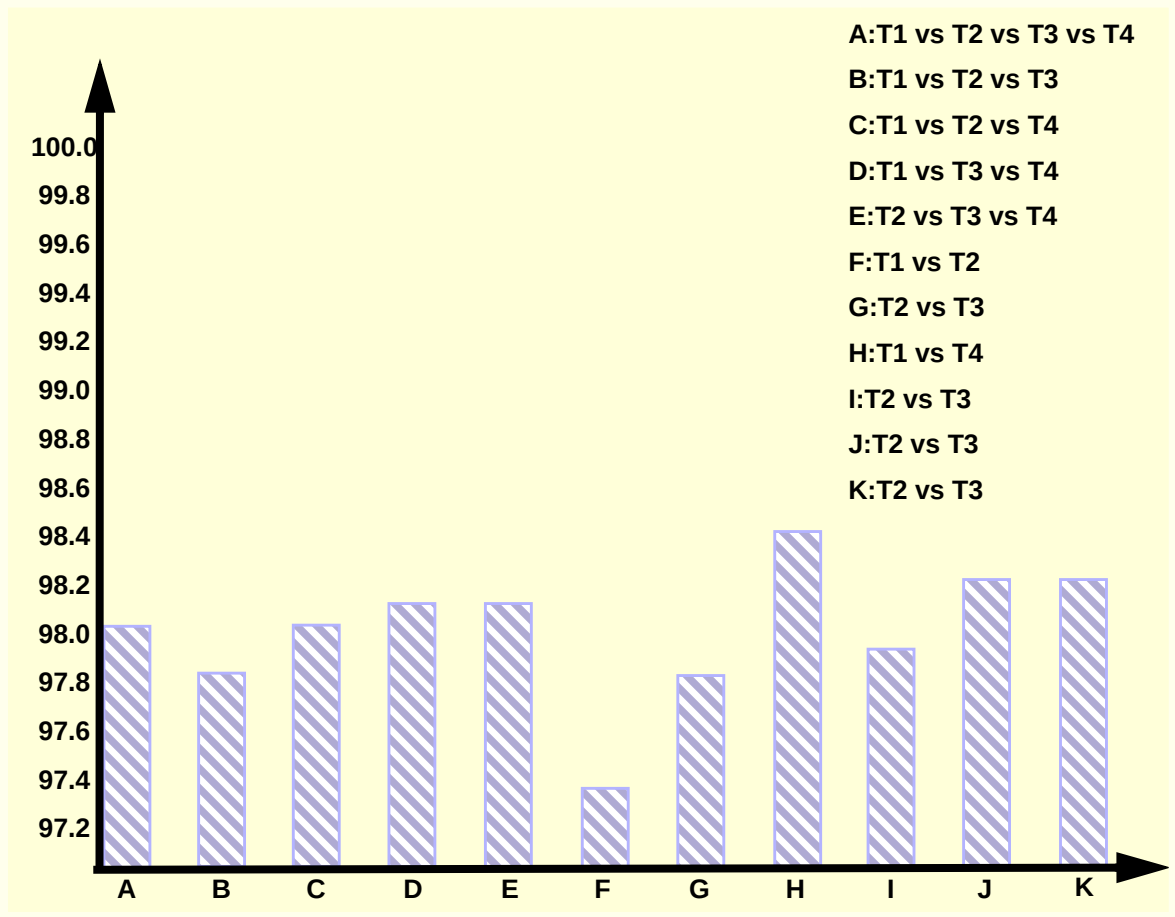


Figure 17. Inter-transcriber agreement on a cross-validation conversation sw4928

Table 5. Error modalities for the cross-validation conversation sw4928

| Modality | Number of words | Inter-transcriber agreement | % Agreement |
|---|---|---|---|
| Normal words | 1030 | 1020 | 99 |
| Silences | 25 | 25 | 100 |
| Laughter words | 2 | 2 | 100 |
| Partial words | 25 | 15 | 60 |
| Proper Nouns | 22 | 20 | 91 |

## Error rates after Word Alignment review

Once the transcription phase of the project was completed, we have proceeded to the next phase: manual review of automatic word alignments generated from a forced alignment using the new transcriptions. The plan was to correct only gross errors found in automatic alignments, particularly those occurring around noise and/or laughter events. This process has not only helped us to offer word alignment data to the speech research community but has also served as a retroactive transcription correction procedure.

### WER estimates based on validator feedback

The feedback regarding transcription errors from the validators reviewing word alignments serves as a good WER estimate for the new transcriptions. The validators get to review around 50-75% of the transcriptions while working on word alignments. Hence their feedback regarding transcription errors can serve as a reliable WER estimate. The error estimates based on such a feedback for 6 conversations is shown in Table 6. A manual review of these conversations has shown the WER estimates to be true. One

Table 6. WER estimates based on validator feedback

| Conversation | Number of words | errors | WER | Modality |
|---|---|---|---|---|
| sw2015 | 1310 | 1 | 0.08 | 1 insertion |
| sw2051 | 1655 | 2 | 0.12 | 1 insertion, 1deletion |
| sw2089 | 2430 | 6 | 0.25 | 6 deletions |
| sw2130 | 1425 | 8 | 0.56 | 7 insertions, 1 deletion |
| sw2171 | 1907 | 2 | 0.10 | 2 deletions |
| sw2125 | 1323 | 2 | 0.15 | 2 deletions |

important observation regarding the error modalities is the absence of substitution errors. A review of WER estimates for 100 more conversations has shown that the WER is stable in the 0.5-1.0% range. Substitution errors account for around 1/10th of the errors. This is is a positive sign and highlights the importance of a stringent QC process to eliminate such errors. The insertion and deletion errors are also mainly concerned with hesitations and repetition words. The review has also shown that the transcriptions are very consistent with respect to non-speech sounds like static and laughter.

## Lexicon Development

It was very important that a corresponding lexicon also be developed in tandem with the transcriptions. A significant amount of time during this project was also devoted to making additions to the lexicon. The conversations were first transcribed by the validators and then modifications made to the segments and transcriptions based on quality control scripts. After this, all words that occurred in the transcriptions but were still absent in the lexicon were flagged by a script along with the utterance identification numbers. Each of these utterances was manually reviewed and a set of words (along with their pronunciations) were proposed as new additions to the lexicon. This set of proposed additions was then reviewed by a group of senior Ph.D. students and the final list released based on their comments. Scripts were also written to make sure that all the phones in the lexicon were contained in the standard phone set (at times, for various reasons, phones outside our phone set can creep in without this check) [40]. Regarding issues such as hyphenated words, we closely followed the conventions of the Merriam-Webster

dictionary [41]. The lexicon also contains pronunciations for partial words and also has common alternate pronunciations.

# CHAPTER V

# EXPERIMENTS AND RESULTS

**N-Gram Statistical Analysis**

**Switched speaker statistics**

**OOV Analysis**

**Disfluency Analysis**

# CHAPTER VI

# SUMMARY AND FUTURE WORK

**Conclusions**

**What we learned from this experience**

**Future Work**

# CHAPTER VII

# REFERENCES


[1] Diana Jecker, “Benchmark Tests: Speech recognition”, *http://www.zdnet.com/pcmag/stories/reviews/0,6755,2385302,00.html,* PC Magazine, November 1999.

[2] P.J. Price and J. Picone, “Automatic Speech Recognition: Better Than Text?” presented at the AAAS Annual Meeting and Science Innovation Exposition, Washington, D.C., USA, February 2000.

[3] A. Martin, M. Przybocki, J. Fiscus, D. Pallet, “The 2000 NIST evaluation for Recognition of Conversational Speech over the Telephone”, presented at the Speech Transcription Workshop, Maryland, USA, May 2000.

[4] L.R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1993.

[5] NIST Spoken Language Technology Evaluations, “*http://www.itl.nist.gov/iaui/894.01/test.html*.

[6] P.J. Price, W.M. Fisher, J. Bernstein and 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.

[7] 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.

[8] 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.

[9] J. Picone, “Signal Modeling Techniques in Speech Recognition”, *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215-1247, September 1993.

[10] C.R. Jankowski, H. Hoang-Doan and L.P. Lippmann, “*A Comparison of Signal Processing Front Ends for Automatic Word Recognition*,” IEEE Transactions on Speech and Audio Processing, vol. 3, no. 4, pp. 286-292, July 1995.

[11] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 29, No. 2, pp. 254-272, April 1981.

[12] V. Mantha, R. Duncan, Y. Wu, J. Zhao, A. Ganapathiraju and J. Picone, “Implementation and Analysis of Speech Recognition Front-Ends”, *Proceedings of the IEEE Southeastcon*, pp. 32-35, Lexington, Kentucky, USA, March 1999.

[13] J. Picone, “Continuous Speech Recognition Using Hidden Markov Models,” *IEEE Acoustics, Speech, and Signal Processing Magazine*, vol. 7, no. 3, pp. 26-41, July 1990.

[14] X. Zhang and Y. Wu, “HMM Training”, *ISIP Fall’98 Seminar Series*, available at http://www.isip.msstate.edu/publications/seminars/isip_weekly/1998/hmm_training/index.html, December 18, 1998.

[15] J.Zhao, “Language Modeling: Integrating Syntactic Constraints”, *ISIP SRSTW00*, available at “http://www.isip.msstate.edu/conferences/srstw00/program/session_08/language_modeling/index.html”, Mississippi State, May 2000.

[16] N. Deshmukh, A. Ganapathiraju, J.Picone, “Hierarchical Search for Large Vocabulary Conversational Speech Recognition”, *IEEE Signal Processing Magazine*, no. 5, pp. 84-107, Sept. 1999.

[17] K. Lee, “Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition”, *IEEE transactions on Acoustics, Speech, and Signal Processing*, pp.599-609, April 1990.

[18] S. Young, “Large Vocabulary Continuous Speech Recognition: a Review”, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 3-28, Snowbird, Utah, USA, December 1995.

[19] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D. Pallett, and N.L. Dahlgren, The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM, available at "*http://www.ldc.upenn.edu/Catalog/docs/TIMIT.html*".

[20] B. Wheatley, G. Doddington, C. Hemphill, J. Godfrey, E.C. Holliman, J. McDaniel, and D. Fisher, "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 533-536, San Francisco, California, USA, March 1992.

[21] 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.htm*l, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, USA, December 1995.

[22] The NIST Spoken Natural Language Processing Group, at *http://www.itl.nist.gov/iaui/894.01/*

[23] 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.

[24] proper noun experiments

[25] disfluency

[26] CLSP-ICSI

[27] 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.

[28] 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, Mississippi State, Mississippi, August 1998.

[29] “Transcription Conventions”, available at *http://www.ldc.upenn.edu/kkarins/convs.paul.h5e.html*.

[30] 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/projects/switchboard/doc/transcription_guidelines*, Institute for Signal and Information Processing, Mississippi State University, July 1998.

[31] N. Deshmukh, J. Hamaker, 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, Mississippi State, Mississippi, USA, August 1998.

[32] R. P. Lippman, “Speech Recognition by Machines and Humans”, *Speech Communication*, Vol 22, pp. 1-15, July 1997.

[33] N. Deshmukh, R. J. Duncan, A. Ganapathiraju, J. Picone, “Benchmarking Human Performance for Continuous Speech Recognition”, *Proceedings of the DARPA Speech Recognition Workshop*, pp. 129-134 Morgan Kaufman Publishers, Harriman, NY, 1996.

[34] J. Hamaker, N. Deshmukh, A. Ganapathiraju and J. Picone, “Improved Monosyllabic Word Modeling on SWITCHBOARD,” Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, August 15, 1998.

[35] J. Picone, “Fundamentals of Speech Recognition: A Short Course,” *http://www.isip.msstate.edu/resources/courses/isip_0000/lecture_notes.pdf*, Institute for Signal and Information Processing, Mississippi State University, May 1996.

[36] J. Hamaker, N. Deshmukh, A. Ganapathiraju and J. Picone, “Improved Monosyllabic Word Modeling on SWITCHBOARD,” Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, November 15, 1998.

[37] V. Mantha, J. Hamaker, N. Deshmukh, A. Ganapathiraju and J. Picone, “Improved Monosyllabic Word Modeling on SWITCHBOARD,” Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, February 15, 1999.

[38] A. Ganapathiraju, N. Deshmukh, V. Mantha and J. Picone, “An Internet-Based Public Domain Speech-to-Text Toolkit”, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, May 15, 1999.

[39] SCLITE

[40] J. Picone, et. al., “Switchboard Resources,” *http://www.isip.msstate.edu/projects/switchboard/doc/education/*, Institute for Signal and Information Processing, Mississippi State University, July 1998.

[41] “Merriam-Webster Online Dictionary”, *http://www.m-w.com*, Merriam-Webster Incorporated, 2000.