AN INTRODUCTION TO THE DIAGNOSTIC EVALUATION OF SWITCHBOARD-CORPUS AUTOMATIC SPEECH RECOGNITION SYSTEMS

Steven Greenberg, Shuangyu Chang and Joy Hollenback

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704

ABSTRACT

A diagnostic evaluation of eight Switchboard-corpus recognition (and six forced-alignment) systems was conducted in order to ascertain whether the associated error patterns can be traced to a specific set of factors. Each recognition system's output was converted to a common format and scored relative to a reference transcript derived from phonetically hand-labeled data (comprising fifty-four minutes of material from several hundred speakers). This reference material was analyzed with respect to several dozen acoustic, linguistic and speaker characteristics, which in turn, were correlated with the recognition-error patterns via a decision-tree analysis. The decision trees indicate that the most consistent factors associated with superior recognition performance pertain to accurate classification of phonetic segments and features. These results suggest that future-generation recognition systems would benefit from improving the acoustic models used for phonetic classification, as well as the pronunciation models involved in lexical matching.

1. INTRODUCTION

The architecture of large-vocabulary speech recognition systems is becoming ever more complex and sophisticated as the demand for enhanced performance and reliability increases. This technological sophistication makes it increasingly difficult to understand a system's underlying architecture, thus stymying efforts devoted to innovation through a principled understanding of why speech recognition systems do not always work as well as they should.

The present study represents an *initial* effort to disect the functional architecture of large-vocabulary speech recognition systems used in the annual NIST-sponsored Switchboard Corpus evaluation. The Switchboard corpus [2] contains hundreds of telephone dialogues of five-to-ten-minute duration between speakers representing a broad cross-section of American society and has been used in recent years (in tandem with the Call Home and Broadcast News corpora) to assess the state of automatic speech recognition (ASR). Switchboard is unique among the large-vocabulary corpora in having a substantial amount of material that has been phonetically labeled and segmented by linguistically trained individuals (Switchboard Transcription Project - http://www.icsi.berekeley.edu/real/stp [3] [6]) and thus provides a crucial set of "reference" materials with which to assess and evaluate the phonetic and lexical classification capabilities of current-generation ASR systems.

This paper focuses on the methods used to evaluate the Switchboard recognition systems, as well as on a few key macroscopic analyses of the diagnostic material. A second paper [5] describes the full spectrum of analyses performed on the Switchboard evaluation material.

2. CORPUS MATERIALS

The evaluation was performed on a fifty-four-minute, phonetically annotated subset of the Switchboard corpus (http://www.icsi.berkeley.edu/real/phoneval). The material had previously been manually segmented at the syllabic and lexical levels and was segmented into phonetic segments using an automatic procedure trained on seventy-two minutes of hand-segmented data from the Switchboard corpus [3]. Approximately 1% of the segmentations were manually adjusted. The output of the automatic segmentation (http://www.icsi.berkeley.edu/real/phoneval) is comparable in reliability to the hand-segmented portion of the corpus.

3. EVALUATION FORMAT

Eight separate sites participated in the evaluation - AT&T, BBN, Cambridge University (CU), Dragon Systems (DRAG), Johns Hopkins University (JHU), Mississippi State University (MSU), SRI International and the University of Washington (UW). Each site was asked to submit two different sets of material:

(1) the word and phonetic-segment output of the recognition system used for the competitive (i.e., non-diagnostic) portion of Switchboard, and

(2) the word and phone-level output of forced-alignments associated with the same material.

The forced-alignments (provided by six of the eight sites) were used to compare the ASR systems' phonetic classification with and without knowledge of the lexicon (cf. Figures 4 and 5).

In order to score the submissions in terms of phone-segments and words correct, as well as perform detailed analyses of the error patterns, it was necessary to convert the submissions into a common format. This required that:

(1) each site's phonetic symbol set be mapped onto a common reference similar to that used to phonetically annotate the Switchboard corpus (STP). Care was taken to insure that the mapping was conservative in order that a site not be penalized for using a symbol set distinct from STP. In addition, phonetic symbols not contained in a site's inventory were mapped to the more fine-grained STP phone set (http:www.icsi.berkeley.edu/real/phoneval).

(2) A reference set of materials at the word, syllable and phone levels was created in order to score the material submitted. This reference material included:

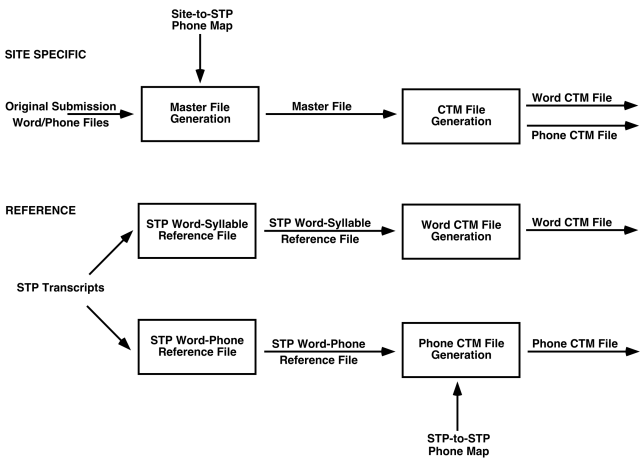

Figure 1: The initial phase of the diagnostic evaluation. Materials submitted by each site are converted into a format designed for scoring (CTM files) relative to the reference transcript (at the phonetic, syllable and word level).

(a) word-to-phone mapping
(b) syllable-to-phone mapping
(c) word-to-syllable mapping
(d) time points for the phones and words in the reference materials

(3) time-mediated synchronization of the phone and word output of the submission material with that of the reference set.

The conversion process (Figure 1) was required in order that the submissions be scored at the word and phonetic-segment levels using SC-Lite, a program developed at the National Institute of Standards and Technology (NIST) to score competitive ASR evaluation submissions.

4. SCORING THE RECOGNITION SYSTEMS

SC-Lite scores each word (and phone) in terms of being correct or not, as well as designating the error as one of three types - a substitution (i.e., *a* -> *b*), an insertion (*a* -> *a*+*b*) or a deletion (*a* -> *ø*). A fourth category, *null*, occurs when the error can not be clearly associated with one of the other three categories (and usually implies that the error is due to some form of formatting discrepancy). A sample, composite (i.e., both phone and word) output from SC-Lite illustrates the scoring method (Table 1).

ID	REF WD	HYP WD	WS	RB	HB	RP	HP	PS	RB	HB
2	ONE	ONE	C	5.70	5.70	W	W	C	5.71	5.70
0						AH	AH	C	5.80	5.81
4						N	N	C	5.88	5.87
0	OUGHT	****	D	5.90		AO	AO	C	5.92	5.90
B						T	D	S	6.06	6.07
0	EITHER	AUDITOR	S	6.10	5.90	IY	IH	S	6.10	6.12
0						DH	T	S	6.30	6.27
1						ER	ER	C	6.34	6.35
1	TO	TO	C	6.43	6.45	T	T	C	6.44	6.45
B						AX	AX	C	6.52	6.52

Table 1: Sample, composite output from SC-Lite showing the scoring method at the word and phone levels. ID (2040-B-0011B) pertains to the entire word sequence, REF WD is the correct word, HYP WD is the recognizer word output, WS is the word score (C = correct, D = deletion, S = substitution), RB is the beginning time (in seconds) of the reference unit (word or phone), HB is the beginning time of the recognizer output, RP is the "correct" phone, HP is the phone output of the recognizer, PS is the phone score.

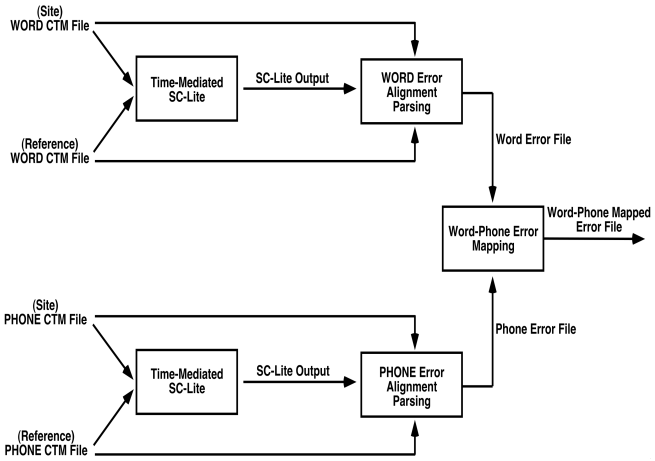

Figure 2: The evaluation's second phase involves time-mediated scoring of both the word- and phone-level output of the recognition and forced-alignment materials. The scored output is used to compile summary tables ("big lists") depicted in Figure 8.

For both phone- and word-scoring it was necessary to develop a method enabling each segment (either word or phone) in the submission to be unambiguously associated with a corresponding symbol (or set of symbols) in the reference material. This was accomplished by using time-mediated boundaries as synchronizing delimiters. Because the word and phone segmentation of the submission materials often deviate from those of the STP-based reference materials an algorithm was developed to minimize the time-alignment discrepancy (parameters 14 and 32 in Table 2).

Files (in NIST's CTM format) were generated for each site's

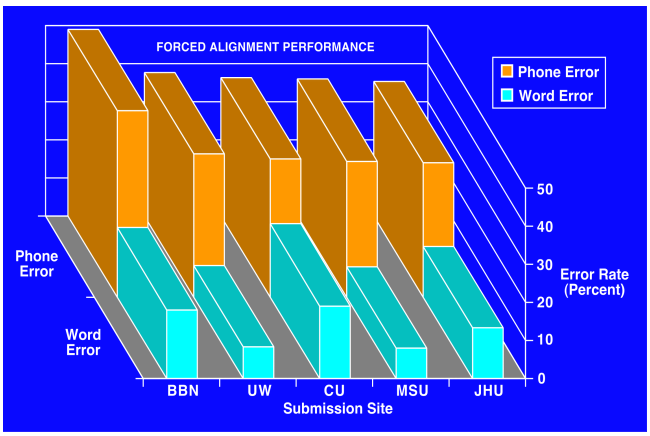

Figure 3: A comparison of the word and phonetic-segment error for the forced-alignment component of the diagnostic Switchboard evaluation for five of the six sites providing this material. The data for SRI are not included due to a mismatch between the lexical representation of their material and that of the reference data. Two sites, Dragon and AT&T, did not provide sufficient data to include in the current analysis.

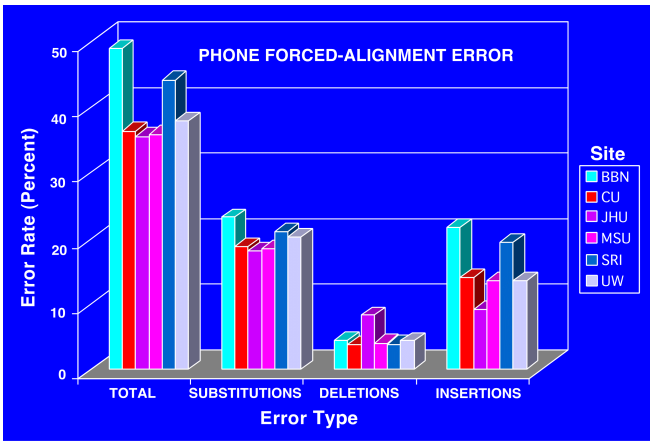

Figure 4: Phonetic-segment errors in the forced-alignment component of the Switchboard diagnostic evaluation for the six participating sites. See text for a description of the site-name code.

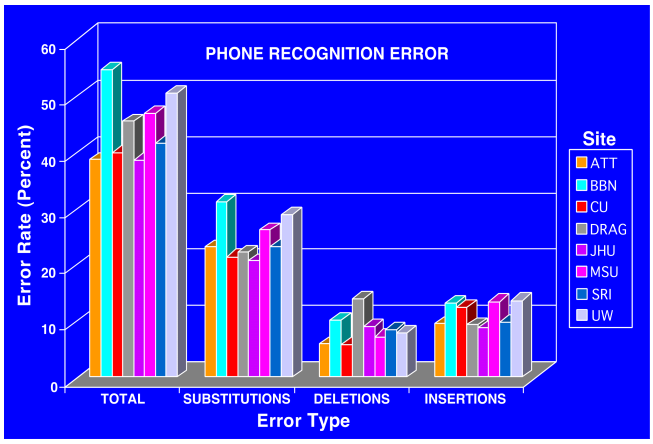

Figure 5: The percentage of phone errors for the recognition component of the Switchboard diagnostic evaluation. Data are from all eight participating sites.

submission (separate files for recognition and forced alignment) and this material processed along with the CTM files associated with the word- and phone-level reference material (Figure 2). The resulting output (Table 1) was used as the basis for generating the data contained in the summary tables ("big lists") described in Section 5.

5. WORD AND PHONE ERROR PATTERNS

Error patterns were computed for both the forced-alignment and recognition submissions. In forced-alignment classification the sequence of words is provided by the word-level transcript - hence this process is mainly a matter of labeling phonetic segments and delineating the segment boundaries. The word errors observed in forced alignment (Figure 3) are usually the consequence of misalignment with the reference material, and are typically of much smaller magnitude than observed in normal recognition. What is of interest is the large number of phone-classification errors observed in the forced-alignment submissions, ranging between 35 and 49% (Figures 3 and 4). Although the error rate is less than that associated with full recognition (39-55% - Figures 5 and 7) the difference in performance between the two conditions is much smaller than anticipated, suggesting that the ASR systems may not be optimized for recognizing phonetic segments.

In the forced-alignment material, there is a relatively high proportion of phone insertions and a disproportionately small number of deletions compared to phone classification in normal recognition. This difference in phone-error pattern is probably due to using word transcripts as the basis for generating hypotheses concerning likely sequences of phones. A fair proportion of "canonical" phone segments are unrealized (i.e., deleted) in spontaneous corpora such as Switchboard, particularly in syllable coda position [4]. The forced alignment phonetic classification may be overly bound to the word transcript and as a consequence "tries too hard" to find phonetic segments where they don't actually occur.

The word error rate for normal recognition systems ranges between 27 and 43%, about 50% higher than that observed for the competitive portion of the evaluation [8]. The higher error rate is probably due to several factors. First, the diagnostic component of the evaluation contains relatively short utterances (mean duration = 4.76 sec) from hundreds of different speakers. In contrast, the competitive evaluation [8] is composed of complete dialogues lasting ca. five minutes and produced by only forty different speakers. Most (if not all) of the recognition systems normally use

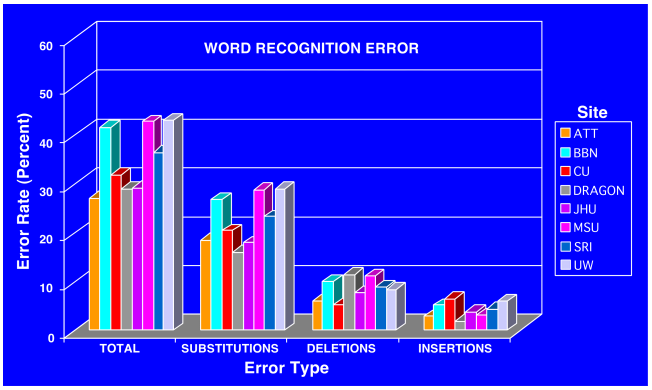

Figure 6: The percentage of word errors for the recognition component of the Switchboard diagnostic evaluation, subcategorized by error type.

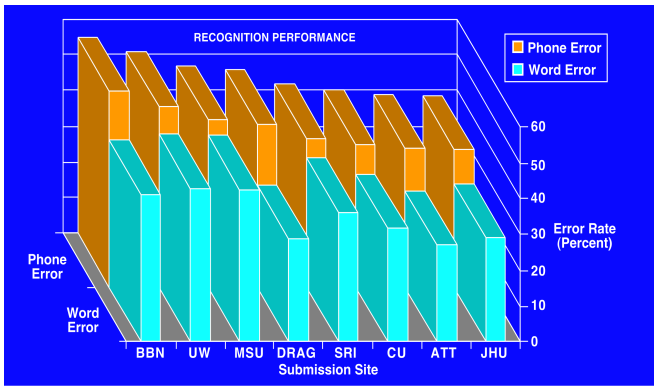

Figure 7: A comparison of the word and phonetic-segment error for the recognition component of the diagnostic Switchboard evaluation for all eight participating sites.

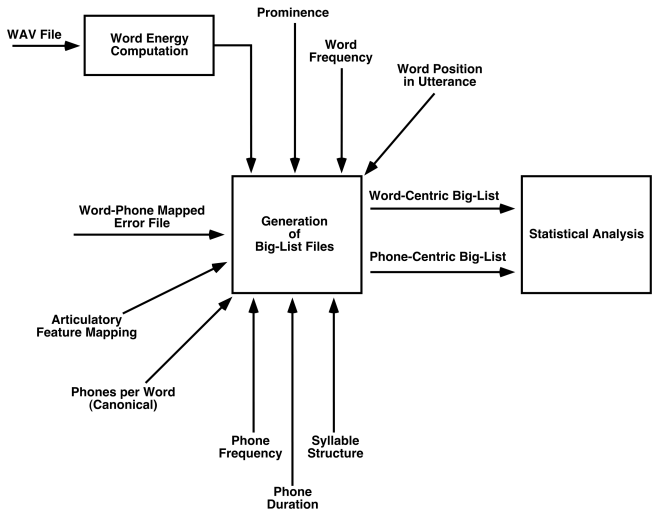

Figure 8: Phase three of the analysis consists of computing several dozen parameters associated with the phone- and word-level representations of the speech signal and compiling these into summary tables ("big lists"). Table 2 contains a complete list of the parameters computed.

some form of speaker adaptation, which works most effectively over long spans of speech. Short utterances, such as those used in the diagnostic evaluation, are likely to mitigate the beneficial effect of speaker adaptation.

Figure 7 illustrates the relationship between phone- and word-error magnitude across submission sites. The correlation between the two *(r)* is 0.78, suggesting that word recognition may largely depend on the accuracy of recognition at the phonetic-segment level (cf. Section 6 for further discussion). Certain sites, such as AT&T and Dragon, deviate from this pattern in that their data contain a lower word error than would be expected based solely on performance at the phonetic-segment level. These systems may possess extremely good pronunciation models that partially compensate for the relative deficiencies of phone classification.

6. DECISION-TREE ANALYSIS OF ERRORS

In order to gain further insight into the factors governing word errors in recognition performance the STP-based, reference component of the Switchboard corpus was analyzed with respect to ca. forty separate parameters pertaining to speaker, linguistic and acoustic properties of the speech materials (Figure 8), including energy level, duration, stress pattern, syllable structure, speaking rate and so on (cf. Table 2 for a complete list of parameters).

Because there are so many different parameters to correlate with word and phone recognition performance the analysis in the current study focuses on decision trees [9] as a means of identifying the most important parameters associated with word-error rate across sites. The error data were partitioned into four separate domains:

(1) Substitutions versus all other data (both correct and incorrect)
(2) Deletions versus all other data (both correct and incorrect)
(3) Substitution versus deletions (i.e., excluding words correctly recognized)
(4) Substitutions versus insertions (i.e., excluding words correctly recognized)

UTTERANCE LEVEL PARAMETERS	
1	**Utterance ID**
2	**Number of Words in Utterance**
3	**Utterance Duration**
4	**Utterance Energy (Abnormally Low or High Amplitude)**
5	**Utterance Difficulty (Very Easy, Easy, Medium, Hard, Very Hard)**
6	**Speaking Rate - Syllables per Second**
7	**Speaking Rate - Acoustic Measure (MRATE)**
LEXICAL LEVEL PARAMETERS	
8	**Word Error Type - Substitution, Deletion, Insertion, Null**
9	**Word Error Type Context (Preceding, Following)**
10	**Word (Unigram) Frequency (in Switchboard Corpus)**
11	**Position of the Word in the Utterance**
12	**Word Duration (Reference and Hypothesized)**
13	**Word Energy**
14	**Temporal Alignment Between Reference and Hypothesized Word**
15	**Lexical Compound Status (Part of a Compound or Not)**
16	**Prosodic Prominence (Maximum and Average Stress)**
17	**Prosodic Context -Maximum/Average Stress (Prec/Following)**
18	**Occurence of Non-Speech (Before, After)**
19	**Number of Syllables in Word (Canonical and Actual)**
20	**Syllable Structure (CVC, CV, etc. - Canonical and Actual)**
21	**Number of Phones in Word (Canonical and Actual)**
22	**Number of Phones Incorrect in the Word**
23	**Type and Number of Phone Errors in Word (Sub, Del, Ins, Null)**
24	**Phonetic Feature Distance Between Hypothesized/Reference Word**
PHONE LEVEL PARAMETERS	
25	**Phone ID (Reference and Hypothesized)**
26	**Phone Duration (Reference and Hypothesized)**
27	**Phone Position within the Word**
28	**Phone Frequency (Switchboard Transcription Corpus)**
29	**Phone Energy**
30	**Phone Error Type (Substitution, Deletion, Insertion, Null)**
31	**Phone Error Context (Preceding, Following Phone)**
32	**Temporal Alignment Between Reference and Hypothesized Phone**
PHONETIC FEATURE LEVEL PARAMETERS	
33	**Manner of Articulation**
34	**Place of Articulation**
35	**Front-Back (Vocalic)**
36	**Voicing**
37	**Lip Rounding**
38	**Cumulative Phonetic Feature Distance**
SPEAKER CHARACTERISTICS	
39	**Dialect Region**
40	**Gender**
41	**Recognition Difficulty (Very Easy, Easy, Medium, Hard, Very Hard)**
42	**Speaking Rate - Syllables per Second and Acoustic (MRATE)**

Table 2: A list of the speaker, utterance, linguistic (prosodic, lexical, phonetic) and acoustic characteristics computed for the diagnostic component of the Switchboard evaluation, the output of which was compiled into summary tables (big lists) for each submission.

ANALYSIS	NODE	ATT	BBN	CU	DRAGON	JHU	MSU	SRI	UW
SUBSTITUTIONS versus all else	1	PHNSUB	PSTWDER	PREWDER	PHNSUB	PHNSUB	PHNSUB	PSTWDER	PHNSUB
	2	WDFREQ	PREWDER	PHNSUB	PREWDER	WDFREQ	AFDIST	PHNSUB	WDFREQ
	3	AFDIST	AFDIST	WDFREQ	PSTWDER	CANSYL#	HYPDUR	AFDIST	AFDIST
	4	BEGINOFF	PHNSUB				PSTWDER	WDFREQ	PSTWDER
DELETIONS versus all else	1	PHNCOR	AFDIST	AFDIST	PHNCOR	AFDIST	AFDIST	PHNCOR	AFDIST
	2	AFDIST	REFDUR	PHNINS	PREWDER	PREWDER	PHNCOR	PHNSUB	REFDUR
	3	PSTWDER	WDENGY	PHNCOR	PHNSUB	REFDUR	PHNINS	PHNINS	PREWDER
	4		PREWDER				PHNSUB	WDFREQ	
SUBSTITUTIONS versus DELETIONS	1	REFDUR	PHNSUB	HYPDUR	REFDUR	REFDUR	REFDUR	REFDUR	REFDUR
	2	PHNSUB	PHNCOR	PHNSUB	PHNSUB	PHNSUB	PHNSUB	WDENGY	PHNSUB
	3	WDENGY	PHNINS	AFDIST	AFDIST	PHNCOR	AFDIST	PHNSUB	PHNINS
	4	PSTWDER		PHNCOR	PHNINS	PHNINS	PHNCOR	WDFREQ	PHNCOR
SUBSTITUTIONS versus INSERTIONS	1	HYPDUR	HYPDUR	HYPDUR	HYPDUR	HYPDUR	PHNFREQ	HYPDUR	HYPDUR
	2	AFDIST	PHNSUB	PHNFREQ		PHNSUB	HYPDUR	PHNSUB	PHNSUB
	3	PREWDER	PHNFREQ	PSTWDER			PREWDER		PHNFREQ
	4		REFDUR				PHNSUB		PHNDEL

Table 3: A summary of the decision-tree analyses performed to identify the primary factors associated with word errors in recognition performance for the eight participating sites. The four most highly utilized (i.e., most important) nodes of the decision tree are shown for four separate analyses (partitioned according to word-error type). The parameters associated with errors are color-coded. Parameters associated with identification of phone segments (PHNSUB, PHNINS, PHNCOR) are indicated in GREEN, those associated with acoustic-phonetic features (AFDIST) are marked in YELLOW, those connected to either word or phonetic segment frequency of occurrence (PHNFREQ, WDFREQ) are indicated in MAGENTA, those pertaining to duration (REFDUR, HYPDUR) are marked in CYAN. Other abbreviations are: BEGINOFF = the temporal disparity between the beginning of a word in the reference transcript and the recognizer output; PREWDER, PSTWDER indicates whether the preceding or following word is also in error, WDENGY = acoustic energy of the word; CANSYL# = the number of syllables contained in the canonical form of the word.

Partitioning of the data was necessary because the underlying cause of an error is likely to depend on whether the error is a substitution (the most common error type), a deletion or an insertion. Each form of error is likely to be associated with a specific constellation of parameters, an assumption that is borne out by the decision-tree analyses (Table 3).

6.1. Substitution Errors (versus All Else)

Two-thirds of the word errors involve substitutions (Figure 6), and it is therefore of interest to identify the parameters associated with this single most important component of recognition performance. For seven of the eight submissions the parameter dominating the decision tree at the highest (or second-highest) node-level is the number of phonetic-segment substitution errors within a word (Table 3). The probability of a word being incorrectly recognized increases significantly when more than (an average of) ca. 1.5 phones are misclassified, and is consistent with the results of an independent analysis performed by Doddington [1]. Other important parameters in the decision trees are the acoustic-articulatory feature distance (AFDIST) between the correct and hypothesized word (which is also related to the probability of correct phone classification), (unigram) frequency of the reference word (WDFREQ), and whether the preceding (PREWDER) or following (PSTWDER) word is incorrectly recognized.

6.2. Deletions (versus All Else)

Deletions account for ca. 25% of the word errors. For all sites, the dominant factor associated with deletion errors pertains to either the number of phonetic segments correctly recognized (PHNCOR) (consistent with the results described in [1]) or the acoustic-articulatory phonetic-feature distance between the reference and hypothesized word. Other important parameters are the number of phone insertions (PHNINS) and substitutions (PHSUB), word frequency and duration of the reference word (REFDUR). The error status of the preceding (PREWDER) and following (PSTWDER) words also appears to play a role.

6.3. Substitution versus Deletion Errors

Additional information concerning the source of word errors can be obtained by analyzing factors distinguishing different types of error. For distinguishing substitution from deletion errors two sets of parameters appear to be most important - phonetic-segment classification (PHNSUB, PHNINS, PHNCOR, AFDIST) and the duration of the reference (REFDUR) and hypothesized (HYPDUR) words.

6.4. Substitution versus Insertion Errors

The duration of the hypothesized word is the most important parameter distinguishing substitution from insertion errors, followed in importance by phonetic segment classification factors (PHNSUB, PHNINS, PHNDEL, AFDIST), frequency of occurrence of the phonetic segments in a word (PHNFREQ) and the error status of the preceding and following words (PREWDER, PSTWDER).

RECOGNITION	ATT	BBN	CU	JHU	MSU	SRI	UW
Converted Submissions	•	•		•	•		•
Word-Level Errors	•	•	•	•	•	•	•
Phone-Level Errors	•	•		•	•		•
Word-Phone Mapping	•	•		•	•		•
Word-Centric Big Lists	•	•	•	•	•	•	•
Phone-Level Big Lists	•	•		•	•		•
Phone-Confusion Matrices	•	•		•	•		•
FORCED-ALIGNMENT							
Converted Submissions		•		•	•		•
Word-Level Errors		•	•	•	•	•	•
Phone-Level Errors		•		•	•		•
Word-Phone Mapping		•		•	•		•
Word-Centric Big Lists		•	•	•	•	•	•
Phone-Level Big Lists		•		•	•		•
Phone-Confusion Matrices		•		•	•		•
SITE PHONE MAPPING	•	•		•	•		•

Table 4: The Phoneval web page (http://www.icsi.berkeley.edu/real/phoneval) contains much (albeit not all) of the materials used in the diagnostic evaluation, as well as summary tables used as the foundation of the error analysis. Listed are the specific set of materials available for each participating site.

6.5. General Trends of the Decision Tree Error Analysis

The most important parameters associated with word-recognition error are those pertaining to the correct identification of a word's phonetic composition (at either the phone or articulatory-acoustic, phonetic-feature level). These results imply that the *single* most effective strategy for reducing word-error rate would be to focus future development on the "front-end" component of recognition systems pertaining to acoustic and phonetic models, as well as on pronunciation models specifying the phonetic composition (and sequence) of lexical units. This conclusion is consistent with the demonstration that error rate can be dramatically reduced by carefully matching the lexical representations (i.e., pronunciation models) to the specific inventory of phonetic segments encountered by the recognition system [7].

The results of the decision-tree analyses are also of interest because of the parameters that are *absent* from the decision trees - prosodic prominence, syllable structure, speaking rate and speaker difficulty. All of these parameters have been suggested as important factors associated with the word-error rate. At first glance it is surprising that they do not appear in the decision-tree analyses. However, it is probably premature to dismiss these non-phonetic parameters as unimportant. Rather, the decision trees suggest that such parameters do not account for word errors "across the board" in the way that acoustic-phonetic factors do. These extra-phonetic characteristics may play an important recognition role for certain speakers or in specific contexts that are not revealed in the decision trees.

7. CONCLUSIONS

Complex systems require sophisticated, multifaceted analyses to characterize their functional architecture and to understand the circumstances under which they fail. The decision-tree analyses represent but one method to account for the pattern of errors observed in recognition of telephone dialogues. Future development of large-vocabulary systems are likely to benefit from focusing on the acoustic-phonetic front-end and on word-pronunciation models as the most efficient means of reducing error rate.

However, decision-tree analyses are most sensitive to factors that pervade the entire corpus and therefore is also necessary to conduct finer-grained analyses of the diagnostic evaluation material in a manner that is not easily accomplished with this technique. Such analyses are described in a separate paper [5] and are available on the Phoneval web site. In addition, an Oracle-based web application, currently under development, will provide extensive data-mining and analysis capabilities for future studies of the Switchboard diagnostic evaluation material (Table 4).

ACKNOWLEDGEMENTS

The authors wish to thank our colleagues at AT&T, BBN, Cambridge University, Dragon Systems, Johns Hopkins University, Mississippi State University, SRI International and the University of Washington for providing the material upon which the diagnostic evaluation of the Switchboard recognition systems is based. We are also grateful to George Doddington, Jon Fiscus, Hollis Fitch, Jack Godfrey, Joe Kupin and Rosaria Silipo for providing valuable assistance with the analyses described, and to Climent Nadeu for comments on an earlier version of this paper. This study was supported by the U.S. Department of Defense.

REFERENCES

1. Doddington, G., "Evidence of differences between lexical and actual phonetic realizations," Presentation at the NIST Speech Transcription Workshop, College Park, MD, May 18, 2000.
2. Godfrey, J. J., Holliman, E. C., and McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pp. 517-520, 1992.
3. Greenberg, S., "The Switchboard Transcription Project," *Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series.* Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD (56 pp.), 1997.
4. Greenberg, S., "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, 29, 159-176, 2000.
5. Greenberg, S., Chang, S, and Hollenback, J., "Linguistic dissection of Switchboard-corpus automatic speech recognition systems," *Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium*, Paris, 2000.
6. Greenberg, S., Hollenback, J. and Ellis, D., "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," *Proc. Int. Conf. Spoken Lang. Proc.*, Philadelphia, S24-27, 1996.
7. McAllaster, D., Gillick, L., Scattone, F. and Newman, M., "Explorations with fabricated data," *Proc. DARPA Workshop Conv. Speech Recog. (Hub-5)*, 1998.
8. Martin, A., Pryzbocki, M., Fiscus, J., and Pallet, D., "The 2000 NIST evaluation for recognition of conversational speech over the telephone," Presentation at the NIST Speech Transcription Workshop, College Park, MD, May 17, 2000.
9. Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.