Analysis of the Errors Produced by the 2004 BBN Speech Recognition System in the DARPA EARS Evaluations

Nicolae Duta, *Member, IEEE*, Richard Schwartz, and John Makhoul, *Fellow, IEEE*

***Abstract*—This paper aims to quantify the main error types the 2004 BBN speech recognition system made in the broadcast news (BN) and conversational telephone speech (CTS) DARPA EARS evaluations. We show that many of the remaining errors occur in clusters rather than isolated, have specific causes, and differ to some extent between the BN and CTS domains. The correctly recognized words are also clustered and are highly correlated with regions where the system produces a single hypothesized choice per word. A statistical analysis of some well-known error causes (out-of-vocabulary words, word fragments, hesitations, and unlikely language constructs) was performed in order to assess their contribution to the overall word error rate (WER). We conclude with a discussion of the lower bound on the WER introduced by the human annotator disagreement.**

***Index Terms*—Error analysis, speech recognition.**

I. Introduction

OVER THE last decade, the large-vocabulary continuous-speech recognition (LVCSR) systems have become more complex and sophisticated in order to respond to the increased demand for accuracy, speed, and reliability [17]. The technological complexity makes it increasingly difficult to understand the recognition systems' behavior and explain why they are not yet working as well as they should [3], [20]. Nevertheless, there has been a continuous effort to analyze the errors incurred in the automatic speech recognition process.

Greenberg *et al.* [5], [6] performed a thorough analysis of the eight systems present in the NIST 2000 Switchboard Corpus evaluation. They used a 54-min subset of the Switchboard corpus which was phonetically annotated with respect to about 40 acoustic, linguistic, and speaker characteristics. The correlation between those data characteristics and the recognition-error patterns was subsequently probed using decision trees. The authors found that the recognition errors were mostly correlated with the number of phonetic-segment substitutions within a word. That is, the probability of a word being incorrectly recognized increased significantly when more than 1.5 phones were misclassified. It was also shown that the speech rate (measured in syllables per second) was highly correlated with the error patterns as well (see also [12]). Utterances slower than three syllables per second or faster than six syllables per second had 50% more recognition errors than utterances within the normal speaking range. Those correlations were found to be consistent over the eight systems analyzed.

Stolcke and Shriberg [21], [22] looked into how speech disfluencies affected the following word predictability within the Switchboard and ATIS corpora. They showed that the language model (LM) transition probabilities were significantly lower at hesitation transitions and that was attributable to both the target word and the word history. It was also suggested that fluent transitions in sentences with a hesitation elsewhere were significantly more likely to involve unmodeled n-grams than transition in fluent sentences. Based on the findings above, the authors listed disfluencies as "one of the factors contributing to the poor performance of the automatic speech recognizers" although they did not show explicit statistics for how disfluencies correlate with the recognition errors. They also proposed a language model that predicted disfluencies probabilistically and took hidden disfluency events into account. Although the model locally reduced the word perplexity, it had no impact on the recognition accuracy.

A recent analysis of spontaneous speech recognition errors appeared in Furui *et al.* [3]. It was performed on 510 min of spontaneous Japanese speech, and it introduced a regression model for the recognition accuracy as a function of six signal and speaker attributes: average acoustic frame likelihood, speech rate, word perplexity, out-of-vocabulary (OOV) rate, filled pause rate, and repair rate. The authors found that the recognition accuracy was mostly correlated with the repair rate and OOV rate and to a somewhat lesser extent with the speech rate. They hypothesized that the strong effect on errors of the repair and OOV rates was due to the fact that "a single recognition error caused by a repair or an OOV word triggers secondary errors due to linguistic constraints."

Several other studies (see [1] and the references therein) attempted to model the relationships between some features present in the speech signal and the recognition word error rate (WER) using logistic regression. The regression model was subsequently used to predict the correctness of the recognition hypotheses.

Manuscript received September 30, 2005; revised May 10, 2006. This work was supported by the Defense Advanced Research Projects Agency under its EARS Program. The associate editor coordinating the review of this paper and approving it for publication was Dr. Isabel Trancoso.

N. Duta was with the Speech and Language Processing Department, BBN Technologies, Cambridge, MA 02138 USA. He is now with the Natural Language Understanding Group, Nuance Communications, Burlington, MA 01803 USA (e-mail: nicolae.duta@nuance.com).

R. Schwartz and J. Makhoul are with the Speech and Language Processing Department, BBN Technologies, Cambridge, MA 02138 USA (e-mail: schwartz@bbn.com; makhoul@bbn.com).

Digital Object Identifier 10.1109/TASL.2006.878268

Palmer and Ostendorf [18] proposed a technique to explicitly model the errors in the speech recognizer's output in order to improve the name entity recognition performance in an information extraction task. They computed statistics for the name entities occurring in the Hub-4 Topic Detection and Tracking data and reported that "the percentage of name words that are OOV is an order of magnitude larger than words in other phrase categories."

In May 2002, the Defense Advanced Research Projects Agency (DARPA) started a research program called EARS (Effective, Affordable, Reusable, Speech-to-text) whose major goal was to reduce recognition word error rates for conversational telephone speech (CTS) and broadcast news (BN) down to the 5%–10% range, running in real-time on a single processor [23]. Progress made in the recognition of English was measured each year on a "Progress Test" (kept fixed for the duration of the program and undisclosed to the participating sites) as well as on "Current Tests" which changed each year and were made public after the official evaluation. Evaluation conditions became more difficult each year by imposing runtime limits, automatic segmentation requirements, and broadening the data sources. However, due to technological improvements and increasingly more data available for training,[1] the word error rates decreased from around 30% for BN and 50% for CTS to around 10% and 15%, respectively. As noted in [17], the EARS-evaluated systems have achieved "remarkable convergence across both sites and domains," with the top systems showing no statistically significant difference in performance [8], [9].

After the 2003–2004 EARS workshops, we performed detailed analyses of the errors our system made in both BN and CTS English evaluations. Since the correlation between acoustic properties of the speech data and the recognition errors was previously investigated [1], [5], [6], we mainly focused on how the errors were distributed, whether they occurred independently, and whether they were correlated with some language properties of the data. Our analyses show that many of the remaining errors are not random but have rather specific causes, occur in clusters, and differ to some extent between the BN and the CTS domains. The BN system is mostly challenged by the proper nouns in the news stories and by the utterance end-points; the CTS system is challenged by a combination of speech disfluencies, high speech rate, and word contraction; and both systems make substitution errors on short or (acoustically) similar words.

The goal of this paper is to quantify the frequencies of the most common error types as well as the errors' correlation with challenging speech events like OOVs, word fragments, hesitations, and disfluent speech. In Section IV, we propose a method to easily detect regions of very high (99%) recognition accuracy in the system's output, which amount to at least half of the test data. One can resegment the test set in order to keep fixed the high-accuracy regions produced in the first decoding stage. Subsequently, it may be possible to reduce the decoding time as well as to improve the recognition performance by combining with the results produced using the original segmentation. Finally, Section V explores the human annotator disagreement when transcribing the same audio and its impact on how low a WER can be achieved.

[1]More than 2000 h of acoustic training data and over 1 billion words of language training data (although only a fraction of the language training is annotated speech) are now available for both BN and CTS.

II. Systems, Models, and Data Description

The recognition results reported in this paper were obtained using the BBN RT04 (Rich Text) system fully described in [13], [19]. In brief, the system consists of the following.

1) A phoneme decoder-based speech segmenter.
2) 14 Perceptual Linear Prediction (PLP) [7] derived cepstral coefficient and energy front-end.
3) A two-pass decoder with state-tied mixture (STM) [14] acoustic and 2-gram LM models in the first pass and state-clustered tied-mixture (SCTM) [15] noncrossword acoustic and 3-gram LM models in the second pass in a Viterbi beam search, followed by either N-best list (for BN) or lattice (for CTS) rescoring using SCTM cross-word acoustic models and 4-gram LM.
4) A two-stage decoding process; the first decoding stage uses speaker independent (SI) models while the second stage uses speaker adaptively trained (SAT) models. The adaptation process is done using two feature-space transforms (a speaker-specific heteroscedastic linear discriminant analysis HLDA [11] and a constrained maximum likelihood linear regression (CMLLR) transform [4]) and 2–16 model parameter transforms (maximum likelihood linear regression (MLLR) [10]).

Our BN system runs in $10\times$ real time (RT) while the CTS system is $20 \times$ RT.

We performed the error analysis on the BN Eval03 and Eval04 test sets and the CTS Eval01 and Deval04[2] sets, which were made available by NIST following the official DARPA evaluations [8], [9]. A quantitative description of the four data sets along with the BBN's system accuracy on them is shown in Table I. All test sets are transcribed by NIST/LDC and also include annotated tokens for disfluent speech (word fragments and hesitations).

We used recognition lexicons of 61K (BN) and 57K (CTS) unique words, to which the most frequent 3K word pairs were added as compounded words. The OOV rate attained was quite small: 0.15%–0.7% over the four sets. The language models we used in this study contained 737 million 4-grams (BN) and 435 million 3-grams (CTS) and were trained on 0.5–1.5 billion words.

III. Qualitative Error Analysis

A. Error Types Present in Both BN and CTS

The main error type that is shared by BN and CTS is substitution of short or (acoustically) similar words (see Table II for a few examples). These errors make up 15% to 25% of all errors. In such cases it is hard even for humans to distinguish among different choices based on local information only. Parsing the sentence might help in a few BN instances, although often the

[2]CTS Deval04 consists of both CTS Eval03 and Dev04 test sets.

TABLE I
SUMMARY OF THE BN AND CTS TEST SETS ON WHICH WE PERFORMED THE ERROR ANALYSIS

Test set	Words	Reference sentences	Segmented utterances and average utt. length	OOVs	Optional words	WER
BN Eval03	24790	508	1318 (19)	47 (0.2%)	380 (1.5%)	8.3%
BN Eval04	46576	935	2358 (19)	320 (0.7%)	1063 (2.3%)	14.2%
CTS Eval01	62909	5895	5895 (11)	149 (0.24%)	2649 (4.2%)	20.1%
CTS Deval04	113991	25725	12623 (9)	167 (0.15%)	5571 (4.9%)	17.0%

TABLE II
EXAMPLES OF SUBSTITUTION OF SHORT OR SIMILAR WORDS IN BN AND CTS RECOGNITION

Reference	Hypothesis
americans who STRUGGLE to understand	americans who STRUGGLED to understand
the cause of a fire that GUTTED A nearly	the cause of a fire that GOT INTO nearly
airlines with THE background TO that	airlines with A background OF that
israeli troops MORE THAN sixty	israeli troops **** WITHIN sixty
stories that will be NEWS later today	stories that will be USED later today
have you done THIS CALL before	*have you done THESE CALLS before*
*from hawaii *** EVEN your parents ARE born*	*from hawaii AND THEN your parents WERE born*
with all the PERVERSION and stuff	*with all the CONVERSIONS and stuff*

TABLE III
EXAMPLES OF WORD-SPLITTING ERRORS (THE REFERENCE IS SPELLED AS A SINGLE WORD) IN BN AND CTS RECOGNITION

HAND WRITTEN	WASTE LAND	WORK WEEK	ICE BOX
OFFICE HOLDERS	COUNTER INTELLIGENCE	SWAMP LAND	CO STARS
SPY MASTER	SCHOOL TEACHER	AFTER SHOCK	MID DAY
MULTI MILLION	*CHEESE BURGERS*	*NON SMOKING*	*HANG OUT*
UNDER RATED	*OVER CROWDED*	*AUTO PILOT*	*SECOND HAND*
ROLLER BLADING	*BREAST FEEDING*	*BLACK OUT*	*EASY GOING*

information necessary to select the "right" choice may be spread across several sentences.

There are also three common error types which are less frequent but which might be easier to fix than the previous ones.

1) Word splitting (or joining) into valid words accounts for 2%–3% of all errors (e.g., HANGOUT → HANG OUT and HARD WORKING→ HARDWORKING, see Table III for more examples). Although the number of such instances is relatively low, each occurrence generates two errors (a substitution along with a deletion or insertion). Many of these cases should be considered equivalent in scoring and for each such possibility one can replace the system output by the most frequently used version.
2) Plurals are often misrecognized as "⟨word⟩ is" (e.g., CARRIAGES → CARRIAGE IS). Some of these errors might be solved using sentence parsing information in a post-processing step.
3) Errors due to inconsistent spelling (e.g., OKAY → O.K, BOUTROUS → BOUTROS, TRAVELLING → TRAVELING). In many cases, the reference is incorrect and one needs to be more careful about spelling conventions.

B. BN Specific Errors

The error analysis revealed the following BN specific errors.

1) Errors generated by proper nouns (person names or places) account for about 10%–15% of the errors (see Table IV for a list of name errors made on BN Eval04). These are mostly due to insufficient training (especially LM training) or no training at all (OOVs, e.g., IVANISEVITCH). We found that about three quarters of the OOV words are name entities.[3] A misrecognized name is often split up and causes several errors (e.g., BRASWELL → BROWN AS WELL) with an average of 1.5–2 errors per word. If the lexicon contains names acoustically close but with different spellings, the system may output any of related spellings (e.g., HANSSEN → HANSEN or HANSON). The mistaken names are usually different on each test set and the 10–15 most frequently misrecognized names account for one third of all name-related errors. A possible

[3]That is somewhat lower than Palmer's estimate [18]. The remaining OOV words consists of rare words (e.g., ESTRANGEMENTS), common words preceded by a prefix (e.g., PROSLAVERY, REPUBLICATION), or improvised words (e.g., SCALAWAG).

TABLE IV
ERRORS GENERATED BY PROPER NOUNS ON BN EVAL04

Name	OOV	Instances mis-recognized	Errors generated	Correctly recognized	Recognized as
VAN LEW	N	26	63	2	THEN LOU
SCALAWAGS	Y	19	42	0	
MALVO	Y	18	25	0	MALVEAUX
DRU SJODIN	Y	9	24	0	DREW SHOULD DEAN
MUHAMMAD	N	15	18	2	MOHAMMED
IAN HUNTLEY	N	6	16	6	
CHEVAUX	Y	9	14	0	SHOW BOAT, SHOULD VOTE
JAWAD AL AMERI	Y	4	14	0	
ACCUWEATHER	N	5	13	2	
KARACHI	N	11	13	7	
CULLEN	N	10	10	1	COLLIN,COLIN,COLLINS
Other names		272	533		
Total		404	785 (12%)		

TABLE V
BOUNDARY WORD ERROR RATE COMPARED TO THE TOTAL WER

Test set	BN Eval03	BN Eval04	CTS Eval01	CTS Deval04
Boundary errors	300	887	1635	2881
Total errors	2065	6594	12672	19403
Boundary WER	11.4%	19.4%	20.7%	17.7%
Total WER	8.3%	14.2%	20.1%	17.0%

solution to the name problem is a time-adaptive lexicon and LM update using training data from a time period immediately preceding the test data [16]. However, the update data does not usually contain sufficient training for the name context, so some context sharing with the regular training data may be needed.

2) There are more errors toward the utterance end-points than there are in the center (e.g., the BN Eval04 WER on the first and the last utterance words is 19% versus 13% on other words, see Table V). This could be due to a segmentation problem (the automatic segmentation misses the true sentence boundary) or just to having less context in the language model. However, the CTS system does not produce a higher WER on end-points neither on Eval01 (manually segmented) nor on Deval04 (automatically segmented).

C. CTS Specific Errors

We have found the following CTS specific errors.

1) A significant number of errors occur around speech disfluencies: hesitations, repeats, partially spoken words.[4] In such cases, both the acoustic and the language model may be inaccurate; since many word sequences are unique and have never occurred before, they cannot be adequately modeled by the language model. We performed a cheating experiment where the small (60K 3-grams) test set was added to the full language model, and that especially helped in these situations (it halved the unadapted WER). A few examples of disfluency-related errors are shown in Table VI.

2) Deletion of word sequences. There are multiple instances where sequences of two to four consecutive words are deleted from the system's output (Table VII). We listened to the audio for 17 such cases, and almost every time, the deletion could be attributed to a combination of severe word contraction, very high speech rate, and low volume. Moreover, in many such cases, the reference was not accurate; it described what the speaker intended to say rather than what he/she actually said.

IV. QUANTITATIVE ERROR ANALYSIS

A. Error Clustering

The alignments between the reference and the best hypothesis suggested that about two thirds of the errors do not occur in isolation but rather in groups of two to eight errors (see first row of Table X). Therefore, the errors do not appear to be independent, since under an independence assumption more than 70% of the errors should be isolated (according to a binomial distribution over samples of the same length as the test utterances). Since the errors are not homogeneously distributed throughout the test set (there are regions, e.g., speaker turns or even full shows, with a much higher error rate than the average), we decided to test the error clustering hypothesis by computing local statistics like the probability of an error given short histories of correct/wrong recognitions. We show the error versus correct state transition automaton in Fig. 1 ($\langle s \rangle$ corresponds to the beginning of a sentence while $\langle /s \rangle$ is used to mark the sentence end).[5] One can notice the following.

[4]That does not imply that the average WER measured around disfluencies has to be higher than the overall WER. Many disfluencies may produce no errors, while others may be very costly. We show a quantitative analysis in Section IV-C.

[5]We only show the transition probabilities for the BN Eval04 and the CTS Deval04 sets. The figures corresponding to the remaining two sets are very similar in each domain and were omitted for space reasons. The transition probabilities were computed under the assumption that the hesitation tokens were NOT optional, fact which slightly increased P(Err) for the CTS domain.

TABLE VI
EXAMPLES OF HESITATION-RELATED ERRORS IN CTS RECOGNITION (HYP DENOTES THE REAL SYSTEM OUTPUT; HYPC IS THE OUTPUT OF THE CHEATING EXPERIMENT)

Ref:	like (%HESITATION) the whole HEAVEN'S gate thing WAS IT HEAVEN'S gate I CAN NOT REALLY (-ember)
Hyp:	like ON the whole HEAVENS gate thing FROM THE HEAVENS gate * THING YOU KNOW
HypC:	like ON the whole heaven's gate thing was it heaven's gate * THING YOU KNOW
Ref:	she had a **** HAUNTED HOUSE (%hesitation) there was a BELL that would ***** RING AT a certain
Hyp:	she had a HARD TO HAVE %hesitation there was a BELLOW that would BRING IT TO a certain
HypC:	she had a haunted house AND there was a bell that would ring at a certain

TABLE VII
EXAMPLES OF WORD-SEQUENCE DELETIONS BY THE CTS SYSTEM

Ref:	to (%hesitation) YOU KNOW all my friends are getting married and EVERY ONE IS HAVING babies
Hyp:	to *** **** all my friends are getting married and ***** *** ** ****** babies
Ref:	maybe AT a higher stage OF DEVELOPMENT THAN we are
Hyp:	maybe ** a higher stage ** ********** THAT we are
Ref:	secretary of state IN THE HOSPITAL in THE hospital after undergoing a serious surgery
Hyp:	secretary of state ** *** ******** in A hospital after undergoing a serious surgery

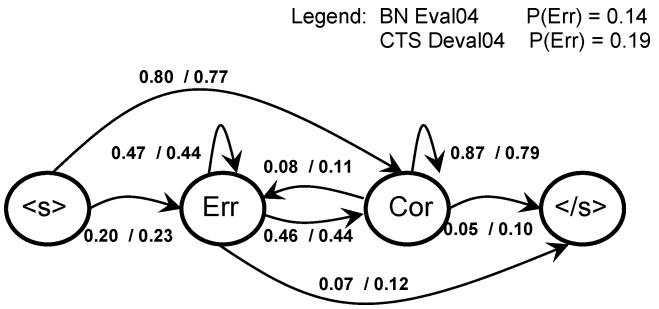


Fig. 1. Transition probabilities between error and correct states for the BN and CTS systems.

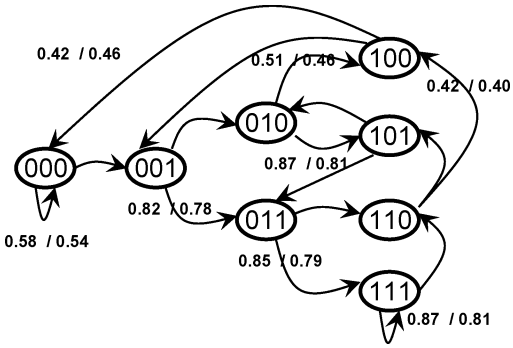


Fig. 2. Transition probabilities for a three word-class (either "correct" or "error") state automaton. A "0" in the state denotes an error output word while a "1" denotes a correct word. Each transition arc is labeled by the probability of observing the right-most word class (either 0 or 1) of the target state given the source state (e.g., the transition from [110] to [101] is labeled by $\mathrm{P}(1|011) = \mathrm{P}(\mathrm{correct}|\mathrm{error}, \mathrm{correct}, \mathrm{correct})$).

1) $\mathrm{P}(\mathrm{Err}|\mathrm{Err}) > 2.5 * \mathrm{P}(\mathrm{Err})$ for both domains, which shows that it is a lot more likely for an error to follow another error than to occur independently of the history.
2) $\mathrm{P}(\mathrm{Err}|\langle \mathrm{s} \rangle)$ and $\mathrm{P}(\mathrm{Err}|\langle /\mathrm{s} \rangle)$[6] (corresponding to errors made on the utterance end-points) are 50% higher than P(Err) for BN, which verifies our direct measurements in Table IV.

A similar automaton corresponding to groups of three adjacent words is shown in Fig. 2 ("0" in a state denotes an error, while "1" denotes a correct word, e.g., "000" represents three consecutive errors). The error clustering trend appears very strong: $\mathrm{P}(\mathrm{Err}|\mathrm{Err}, \mathrm{Err}, \mathrm{Err})$ is 2.5 to 3 times higher than P(Err), and even when the history contains a correct word, one still has a much increased probability of error. As expected, the correctly recognized words are also strongly clustered. However, as long as the most recent word is correct, the remaining history does not matter anymore: $\mathrm{P}(\mathrm{Cor}|\mathrm{Cor}, \mathrm{Cor}, \mathrm{Cor}) = \mathrm{P}(\mathrm{Cor}|\mathrm{Cor}, \mathrm{Err}, \mathrm{Cor}) = \mathrm{P}(\mathrm{Cor}|\mathrm{Cor}) = \mathrm{P}(\mathrm{Cor}) = 1 - \mathrm{P}(\mathrm{Err})$. That is, for correctly recognized words, the third-order Markov model is reduced to a first-order model, while for errors it is still a third-order model.

[6]According to Bayes' law, $\mathrm{P}(\mathrm{Err}|\langle /\mathrm{s} \rangle) = \mathrm{P}(\langle /\mathrm{s} \rangle|\mathrm{Err}) * \mathrm{P}(\mathrm{Err}) / \mathrm{P}(\langle /\mathrm{s} \rangle) = \mathrm{P}(\langle /\mathrm{s} \rangle|\mathrm{Err}) * \mathrm{P}(\mathrm{Err}) / [\mathrm{P}(\langle /\mathrm{s} \rangle|\mathrm{Err}) * \mathrm{P}(\mathrm{Err}) + \mathrm{P}(\langle /\mathrm{s} \rangle|\mathrm{Cor}) * \mathrm{P}(\mathrm{Cor})] = 0.07 * 0.14 / [0.07 * 0.14 + 0.05 * 0.86] = 0.19$.

B. Identifying Clusters of Correctly Recognized Words

Most state-of-the-art LVCSR systems employ some statistical measure to assess the confidence in the system's output. In this section, we propose a simple method for estimating regions of correctly recognized output.

For each test set, we aligned the list of the 100 best hypotheses, and we analyzed the regions that only had a single choice for each word. Fig. 3 shows a single word choice versus multiple word choice automaton computed using the hypotheses generated after the second (speaker adapted) decoding stage. This automaton has a clustering trend similar to that in Fig. 1 on both BN and CTS systems and all four test sets. Given that we are in a single choice region, the probability to remain there is 0.66 while the overall probability of a single choice word

TABLE VIII
RECOGNITION STATISTICS ON THE OPTIONAL TOKENS (HESITATIONS AND WORD FRAGMENTS)

Test set	Optional tokens	Hesitations/word fragments deleted	Optional full words recognized	Optional full words deleted	Optional tokens substituted
BN Eval03	380 (1.5%)	348 (92%)	0	0	32 (8%)
BN Eval04	1063 (2.3%)	876 (82%)	8	9	164 (18%)
CTS Eval01	2649 (4.2%)	1888 (71%)	34 (1.5%)	32 (1.5%)	680 (26%)
CTS Deval04	5571 (4.9%)	4234 (76%)	85 (1.7%)	185 (3.3%)	1057 (19%)

TABLE IX
RECOGNITION STATISTICS ON THE UNLIKELY LANGUAGE CONSTRUCTS

Test set	Unlikely constructs	Correctly recognized	Misrecognized	Total damage (errors)
BN Eval03	1065 (4.3%)	862 (81%)	203 (19%)	392
BN Eval04	2301 (5.0%)	1755 (76%)	546 (24%)	1228
CTS Eval01	2312 (3.7%)	1801 (78%)	511 (22%)	1228
CTS Deval04	4082 (3.6%)	3201 (78%)	881 (22%)	2044

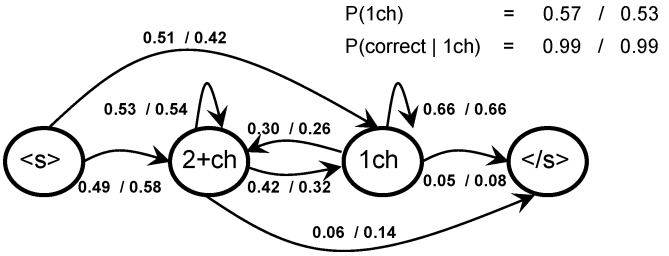


Fig. 3. Transition probabilities between single-choice and multiple-choice states for the BN and CTS systems following the second (speaker adapted) decoding stage.

is only 0.55. At the same time, $\mathrm{P}(\mathrm{Cor}|\mathrm{single\ choice}) = 0.99$. In other words, there is 99% recognition accuracy on the single-choice region (about 55% of all words) of each test set.

Similar results are obtained if the hypotheses generated after the unadapted decoding are used in computing the transition probabilities. The only difference is that P(single choice) is slightly lower: 0.53 for BN and 0.48 for CTS. That is, the regions of high recognition confidence are smaller when the unadapted system output is used (compared to the output generated by the adapted system). We noticed that most of the 1% errors found in the single choice per word regions using the unadapted hypotheses are not fixed after adapted decoding.

If the test set is resegmented at the boundaries of the single word choice regions, it is possible that redecoding only the multiple word choice regions in the subsequent adaptation stages could help in two ways: 1) speed-up the system and 2) improve accuracy by allowing system combination with the results obtained using the original (unadapted) segmentation.

C. Impact of Nonfluent and Nonmodeled Speech on Errors

We have also measured how nonmodeled words (OOVs), word fragments, unintelligible speech (generically marked as "%hesitation" by the human annotators), as well as other forms of nonfluent speech (repeats, fillers, edits)[7] influence the WER. One should first note that the reference tokens marked as word fragments and unintelligible speech are optionally deletable for scoring purposes. That is, one introduces an error if such a token is substituted but not if it is deleted. All optional tokens are considered when computing the total number of reference words by which one normalizes the WER.

As shown in Table VIII, about 1.5%–2.5% (BN) and 4%–5% (CTS) of all reference words are marked as optional, and they are a lot more frequent in CTS than in BN. Very few (<6%) of the optional tokens are actually full words which can be correctly recognized. According to Columns 4–5 of Table VIII, 30%–50% of them are indeed correctly recognized. All other tokens are either nonmodeled word fragments or generic hesitations. Both BN and CTS systems are tuned to delete 70%–90% of the optional tokens in order not to introduce errors. As a consequence, especially our CTS system, avoids producing output on some high rate speech regions and on partially spoken words although those words are not marked as optional in the reference. That agrees with our observation in Section III-C that the CTS system is unbalanced toward deletions.

The recognition statistics for the unlikely language constructs are shown in Table IX. Since we used very large language models, a word pair (word, word history) was not explicitly modeled only 3.5%–5% of the time. In most (75%–80%) of these nonmodeled cases, the system still produced the correct output. However, the mistakes due to unlikely language constructs are very costly: each misrecognized word generates multiple errors (see last column in Table IX).

Table X shows the (per cluster) distribution of the errors generated by the three event types: OOVs, optional tokens, and unlikely language constructs. The count of each event (and its associated error count) was computed for each error cluster length

[7]These events were not explicitly marked as such in our references. However, we considered them to occur in regions that did not contain OOVs but in which our very large LMs had to be backed-off up to a unigram. Given that our LM's bigram hit rate is 98% on fluent (like newspaper) text, there is only a 2% chance that a fluent word pair is not modeled; most remaining pairs are examples of unlikely language constructs.

TABLE X
STATISTICS OF THE ERROR DISTRIBUTION (COUNTS OF THE ERRORS OCCURRING IN CLUSTERS/GROUPS OF LENGTH i) ALONG WITH ERROR CONTRIBUTION FROM OOVS, OPTIONAL TOKENS, AND UNLIKELY LANGUAGE EVENTS. THE FIRST FIGURE IS THE EVENT COUNT, THE SECOND IS THE ASSOCIATED ERROR COUNT (e.g., SECOND COLUMN IN ROW OOVS SHOWS THAT 254 OF THE 1744 ERRORS THAT OCCUR IN GROUPS OF TWO ARE GENERATED BY 131 OOV WORDS FOUND IN 127 TWO-ERROR CLUSTERS)

BN Eval04	**Counts of [target events, errors associated] present in clusters of length i:**								
Target events	1	2	3	4	5	6	7	8	**Total**
All errors	2035	1744	1077	600	315	228	112	112	6594
OOVs	72/72	131/254	60/153	29/92	17/65	6/36	2/14	0/0	320/714
Optional tokens	91/91	31/60	20/51	14/24	0/0	1/6	6/7	0/0	164/248
Unlikely language	182/182	163/308	86/231	45/148	34/130	16/84	10/56	4/32	546/1228

CTS Deval04	**Counts of [target events, errors associated] present in clusters of length i:**								
Target events	1	2	3	4	5	6	7	8	**Total**
All errors	6613	4994	2886	1916	1100	696	371	232	19255
OOVs	38/38	55/110	21/60	20/80	9/45	9/48	2/14	1/8	167/473
Optional tokens	611/611	232/408	100/222	54/148	39/105	13/42	2/14	2/16	1057/1068
Unlikely language	337/337	251/498	116/348	78/300	49/225	27/138	7/42	3/24	881/2044

TABLE XI
STATISTICS OF THE LANGUAGE MODEL EVALUATION ORDER MEASURED ON THE SYSTEM'S OUTPUT (1-BEST HYPOTHESIS) FOR THE ERROR SAND CORRECT REGIONS AS WELL AS ON OOVS, OPTIONAL TOKENS, AND UNLIKELY LANGUAGE CONSTRUCTS. THE EVALUATION ORDER MEASURED ON THE REFERENCE FOR THE ERROR REGIONS IS ALSO SHOWN FOR COMPARISON

Target events	**LM evaluation order (BN Eval04)**					**LM evaluation order (CTS Deval04)**			
	0	1	2	3	4	0	1	2	3
All error words (reference)	4%	12%	23%	23%	31%	9%	6%	12%	73%
All error words (1-best hypothesis)		9%	32%	30%	29%		2%	11%	87%
Correct words		3%	18%	31%	48%		1%	5%	94%
OOV words		24%	37%	23%	16%		11%	17%	72%
Optional token errors		6%	30%	30%	34%		1%	10%	89%
Unlikely language errors		24%	37%	22%	17%		6%	17%	77%

i (i $=$1 to 8). For example, on BN Eval04, 72 isolated errors were generated by OOVs, while 131 OOVs occurred in 127 error clusters of length 2 and therefore produced 254 errors. After manually inspecting the error clusters, it appears that for small values of i (2 to about 4) all the errors in a cluster in which one of the three target events mentioned above occurs, can be attributed to that target event.[8] According to Table X, the unlikely language constructs produce the most damage (2.5 errors per occurrence), followed by OOVs (two errors per occurrence) and by optional tokens (1.5 errors per occurrence). This result confirms the hypothesis in Furui *et al.* [3].

It is also interesting to consider the language model behavior on the error and correct clusters as well as on the three event classes mentioned previously. Before measuring this behavior, we have intuitively assumed that whenever a higher order {3–4} n-gram was not modeled by the LM, the recognition system had to consider a shorter history and back-off the probability until the (target, history) was actually modeled. Tables IX and XI show that is the case most of the time. However, when a word is not modeled by the LM and about 8%–15% of the cases the pair (word, immediate history) is not modeled, the system prefers to use higher order n-grams which acoustically resemble the utterance. That is, instead of using the correct 1-gram, the system uses an incorrect {3–4}-gram. In such cases, a whole neighborhood of the target word is misrecognized and multiple errors are generated. That explains why errors due to OOVs and unlikely language are so costly and often occur in 2–4 word clusters.

[8]We noticed that the long error clusters (some of which span the entire utterance) can rather be attributed to low-quality (very fast, low volume, accented, noisy) speech, so one can consider them outliers.

V. DISCUSSION: ERROR MEASUREMENT

The automatic speech recognition errors are defined by the disagreement between the output of the automatic system and the output of the human recognition (typically called ground truth reference) on the same speech data. We would like to conclude the paper with a discussion of the error rate dependence on the human-made ground truth.

The error measure, called word error rate, is computed as the sum of the errors in each of the three classes (substitutions, insertions, and deletions) and is normalized by the number of reference words. Usually, a single manually generated and

carefully annotated (by two independent transcribers with the disagreements adjudicated by a third person) reference is used as a ground truth. Although transcriptions are done carefully, the references produced by different transcriber teams are not identical.

We present our attempt at quantifying and explaining the annotation differences (for a full statistical analysis see [2]). In 2003, BBN contracted WordWave to transcribe 1700 h of Fisher data [9] to be distributed to the EARS community for CTS acoustic training. In order to measure the quality of the "quick" transcriptions, WordWave was asked to transcribe the CTS Eval03 test set for which a careful transcription was provided by MSU-LDC. After alignment, the WordWave transcription showed 11.5% WER w.r.t. to the MSU-LDC transcription. We randomly picked and listened to 15 of the 144 5-min speaker turns which had multiple transcription differences (343 out of 2511 words) and found the following.

1) In about 30% of the cases, the MSU transcription appeared to be correct, some of the differences may have been due to carelessness or fatigue of the WordWave transcriber.
2) In about 15% of the cases, the WordWave transcription appeared to be correct, we noticed a few differences on words with foreign origin (e.g., "LA RUE GASTRONOMI QUE") as well as some cases where MSU transcribed what the speaker intended to say rather than what he/she actually said.
3) In about 25% of the cases, we could not tell which transcription was correct; much of the speech was not audible and there was true ambiguity in the utterance.
4) About 25% of the cases were different spelling conventions (e.g., UH versus AH).
5) About 10% of the differences are due to incomplete annotations of NOISE or LAUGHTER which each transcriber may mark somewhat randomly if the audio is noisy.

After normalizing the spelling conventions and eliminating the NOISE markings, the real differences between the two transcriptions were around 6%–7% (this figure was later confirmed in [2] on multiple transcription sets). As the speech-to-text WER will soon approach the differences among transcribers, we will have to account for these differences when computing the WER. To overcome this problem, several alternative error measures were introduced in [2].

VI. Conclusion

In this paper, we quantified the main error types still present in a speech recognizer's output and measured their correlation with some language properties of the data. We showed that there are both common and specific error types in BN and CTS. However, the main error types are somewhat different.

1) In comparison with BN data, CTS data contains very few name entities, and even though each name still causes more than one error when misrecognized, the total number of name-related errors is small.
2) The disproportionate percentage of errors that occur at the utterance end points in BN did not occur for CTS. It is unclear at this point whether that is due to the test set segmentation or to a weak LM at sentence boundaries.
3) The large percentage of deletions that occur in CTS shows that the system is tuned to avoid errors in regions of disfluent speech a significant number of which are marked as optional. In this way, the average WER around disfluencies does not become higher than the average WER. However, some disfluencies may generate multiple errors (see Tables VI and X).

The four test sets analyzed were consistent with respect to the error types and frequencies. The only exception was the misrecognition of proper names, which was very much dependent on the time period when the test set was collected. Finally, the error analysis shows that many of the remaining errors are not random but have rather specific causes. The challenge is now how to use this information to reduce the WER. That might be possible by designing different solutions for different error classes, and the detection of possible error, or correct regions might aid in this error class specific process.

References

[1] S. Choularton. Investigating the Acoustic Sources of Speech Recognition Errors. [Online] Available: http://www.ics.mq.edu.au/~stephenc/inter2005.pdf

[2] J. Fiscus and R. Schwartz, "Analysis of scoring and reference transcription ambiguity," presented at the *EARS 2004 Meeting*, Palisades, NY, Nov. 7–10, 2004.

[3] S. Furui, M. Nakamura, T. Ichiba, and K. Iwano, "Why is the recognition of spontaneous speech so hard?," in *Proc. 8th Int. Conf. Text, Speech, Dialogue*, Karlovy Vary, Czech Republic, 2005, pp. 9–22.

[4] M. J. F. Gales, "Maximum-likelihood linear transformation for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, pp. 75–98, 1998.

[5] S. Greenberg, S. Chang, and J. Hollenback, "An introduction to the diagnostic evaluation of the Switchboard-corpus automatic speech recognition systems," in *Proc. NIST Speech Transcription Workshop*, College Park, MD, May 16–19, 2000, [Online] Available: http://www.nist.gov/speech/publications/tw00/pdf/cp2110.pdf..

[6] S. Greenberg and S. Chang, "Linguistic dissection of switchboard-corpus automatic speech recognition systems," in *Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, Paris, France, 2000, pp. 195–202.

[7] H. Hermansky, "Perceptual linear predictive PLP analysis for speech," *J. Acout. Soc. Amer.*, vol. 4, pp. 1738–1752, 1990.

[8] A. Le. Rich transcription 2003 spring speech-to-text evaluation results. presented at *EARS 2003 Meeting*. [Online] Available: http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/rt03s-stt-results-v9.pdf

[9] ——, 2004 fall rich transcription speech-to-text evaluation. presented at *EARS 2004 Meeting*. [Online] Available: http://www.nist.gov/speech/tests/rt/rt2004/fall/rt04f-stt-results-v6b.pdf

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Comput. Speech Lang.*, vol. 9, pp. 171–186, 1995.

[11] S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," in *Proc. IEEE ASRU Workshop*, St. Thomas, U.S. Virgin Islands, 2003, pp. 273–278.

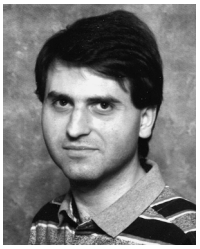
[12] N. Mirghafori, E. Fosler, and N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes," in *Proc. Eurospeech Conf.*, Madrid, Spain, 1995, pp. 491–494.

[13] L. Nguyen, B. Xiang, M. Afify, S. Abdou, S. Matsoukas, R. Schwartz, and J. Makhoul, "The BBN RT04 English broadcast news transcription system," in *Proc. Interspeech Conf.*, Lisbon, Portugal, 2005, pp. 1673–1676.

[14] L. Nguyen and R. Schwartz, "Single-tree method for grammar-directed search," in *Proc. ICASSP Conf.*, Phoenix, AZ, 1999, pp. 613–616.

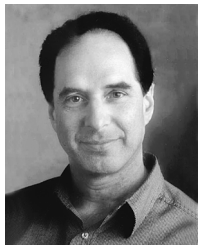
[15] ——, "Efficient 2-pass N-best decoder," in *Proc. Eurospeech Conf.*, vol. I, Rhodos, Greece, 1997, pp. 167–170.

[16] K. Ohtsuki, N. Hiroshima, M. Oku, and A. Imamura, "Unsupervised vocabulary expansion for automatic transcription of broadcast news," in *Proc. ICASSP Conf.*, vol. I, Philadelphia, PA, 2005, pp. 1021–1024.
[17] M. Ostendorf, E. Shriberg, and A. Stolcke, "Human language technology: Opportunities and challenges," in *Proc. ICASSP Conf.*, vol. V, Philadelphia, PA, 2005, pp. 949–953.
[18] D. Palmer and M. Ostendorf, "Improving information extraction by modeling errors in speech recognizer output," in *Proc. Human Language Technology Workshop*, San Diego, CA, 2001, pp. 1–5.
[19] R. Prasad, S. Matsoukas, C. Kao, J. Ma, D. Xu, T. Colthurst, O. Kimball, R. Schwartz, J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre, "The 2004 BBN/LIMSI $20 \times$ RT English conversational telephone speech recognition system," in *Proc. Interspeech Conf.*, Lisbon, Portugal, 2005, pp. 1645–1648.
[20] E. Shriberg, "Spontaneous speech: How people really talk, and why engineers should care," in *Proc. Interspeech Conf.*, Lisbon, Portugal, 2005, pp. 1781–1784.
[21] E. Shriberg and A. Stolcke, "Word predictability after hesitations: A corpus-based study," in *Proc. Int. Conf. Spoken Language Processing*, Philadelphia, PA, 1996, pp. 1868–1871.
[22] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proc. ICASSP*, Atlanta, GA, 1996, pp. 405–408.
[23] C. Wayne, "Effective, affordable, reusable, speech-to-text," presented at the *EARS 2003 Meeting*, Boston, MA, May 19–22, 2003.



Nicolae Duta (M'91) received the B.S. degree in applied mathematics from the University of Bucharest, Bucharest, Romania, in 1991, the D.E.A. degree in statistics from the University of Paris-Sud, Paris, France, in 1992, the M.S. degree in computer science from the University of Iowa, Iowa City, in 1996, and the Ph.D. degree in computer science and engineering from Michigan State University, East Lansing, in 2000.

He is currently a Scientist in the Natural Language Understanding Group, Nuance Communications, Burlington, MA. From 2000 to 2005 he was a Scientist in the Speech and Language Processing department at BBN Technologies, Cambridge, MA. He also held temporary research positions at INRIA-Rocquecourt, France, in 1993 and Siemens Corporate Research, Princeton, NJ, from 1997 to 1999. His current research interests include computer vision, pattern recognition, language understanding, automatic translation, and machine and biological learning.



Richard Schwartz received the S.B. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

He joined BBN Technologies, Cambridge, MA, in 1972 and is currently a Principal Scientist. He specializes in speech recognition, speech synthesis, speech coding, speech enhancement in noise, speaker identification and verification, machine translation, and character recognition.

John Makhoul (F'80) received the B.E. degree from the American University of Beirut, Beirut, Lebanon, the M.Sc. degree from the Ohio State University, Columbus, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, all in electrical engineering.

Since 1970, he has been with BBN Technologies, Cambridge, where he is a Chief Scientist working on various aspects of speech and language processing, including speech recognition, optical character recognition, language understanding, speech-to-speech translation, and human–machine interaction using voice. He is also an Adjunct Professor at Northeastern University, Boston, MA.

Dr. Makhoul has received several IEEE awards, including the IEEE Third Millennium Medal. He is a Fellow of the Acoustical Society of America.