UNDERSTANDING AND IMPROVING SPEECH RECOGNITION PERFORMANCE THROUGH THE USE OF DIAGNOSTIC TOOLS

Ellen Eide, Herbert Gish, Philippe Jeanrenaud, and Angela Mielke

BBN Systems and Technologies
70 Fawcett St. 15/1b
Cambridge, MA 02138

1. INTRODUCTION

The goal of this work is to highlight aspects of an experiment other than the word error rate. When a speech recognition experiment is performed, the word error rate provides no insight into the factors responsible for the recognition errors. We begin this paper by describing an experiment which contrasts the language of conversational speech with that of text from the Wall Street Journal. The remainder of the paper is devoted to the description of a more general approach to performance diagnosis which identifies significant sources of error in a given experiment. The technique is based on the use of binary classification trees; we refer to the results of our analyses as diagnostic trees. Beyond providing understanding, diagnostic trees allow for improvements in the performance of a recognizer through the use of feedback provided by quantifying confidence in the recognition.

2. THE LANGUAGE OF THE SWITCHBOARD AND WSJ CORPORA

The experiment described in this section was designed to highlight the performance attainable on sentences from the Switchboard corpus relative to that achievable on Wall Street Journal sentences. We contrasted the language of the two data sets while removing other ways in which the two databases differ by recollecting in-house speakers reading both sets of transcriptions. Six native English speakers read 100 Switchboard and 50 WSJ sentences using a Sennheiser microphone. Half of the Switchboard sentences were unique to a speaker and the other half were common to all speakers. The utterances were verified and those deemed unacceptable for any reason were discarded.

	Read Switchboard				
	Sub	Del	Ins	Err	Words
Spkr-1-Female-Swbd	16.9	1.1	3.7	21.7	783
Spkr-2-Male-Swbd	27.6	8.6	4.1	40.3	660
Spkr-3-Male-Swbd	19.5	2.0	3.2	24.7	665
Spkr-4-Female-Swbd	23.9	3.1	7.2	34.2	552
Spkr-5-Female-Swbd	17.1	1.6	5.4	24.1	744
Spkr-6-Female-Swbd	16.7	2.1	3.5	22.3	623
Average	20.0	3.0	4.4	27.5	4027

Table 1: Performance on reading Switchboard sentences.

Each set of utterances was decoded using acoustic models built from WSJ training. We used a trigram model built from only the appropriate corpus for decoding each of the two sets of utterances. The word perplexity of the Switchboard test set was 68, while that of the WSJ sentences was 77.

	Read WSJ				
	Sub	Del	Ins	Err	Words
Spkr-1-Female-WSJ	3.7	0.5	2.1	6.3	934
Spkr-2-Male-WSJ	7.9	1.4	1.1	10.4	710
Spkr-3-Male-WSJ	4.7	1.5	0.7	6.9	802
Spkr-4-Female-WSJ	9.5	0.7	2.4	12.6	718
Spkr-5-Female-WSJ	4.1	0.4	1.0	5.5	700
Spkr-6-Female-WSJ	6.2	1.7	1.3	9.2	776
Average	5.9	1.0	1.5	8.4	4640

Table 2: Performance on reading WSJ sentences. The speakers are the same as those in table 1.

Shown in tables 1 and 2 are the respective error rates on the Switchboard and WSJ portions of the experiment. In addition to measuring the overall word error rate, we have looked at error as function of word type and length, as shown in table 3.

Category	WSJ		Read Swbd	
	%	Error	%	Error
Function word	35	13.5	59	30.0
More than 3 phonemes	56	5.0	29	15.2
1,2,or 3 phonemes	44	12.8	71	32.3
1 or 2 phonemes	22	15.5	39	34.0
1 phoneme	3	32.7	8	42.1

Table 3: Breakdown of error according to word category for the read versions of Switchboard and WSJ. Shown in the left column of each set are the percentage of words which fall into the category. In the right column are the error rates within each category.

The error rates on the Switchboard sentences are consistently much higher than on WSJ. The major discrepancy between the two data sets lies in the length of the words; in the Switchboard transcriptions the average word length is 3.0 phonemes, whereas on WSJ the average is 4.1 phonemes. Although the word-level perplexities or average branching factors are comparable, we must branch 35% more often in Switchboard, rendering the phoneme-level perplexity of Switchboard much higher than that of WSJ [1].

In addition to comparing the results of the read Switchboard sentences with performance on WSJ, we have compared the performance on the read and conversational, telephone Switchboard

0-7803-2431-5/95 $4.00

data. A common subset of 39 sentence transcriptions was shared between the test set read for this experiment and the CAIP development test set of the November 1994 LVCSR evaluations. Although this set of 39 sentences is too small to be a reliable result on its own, it is interesting to note that the performance for the read data on this set was 32.9%, and for the conversational telephone data the error was 48.0%.

The abundance of short words in conversational speech is at the heart of the difficulty in recognizing it, accounting for the largest increase in error rate incurred in going from read WSJ to conversational Switchboard. The problem of short words has been addressed through a scheme of pairing short words to form longer compound words; this tactic is discussed more fully in [2].

3. THE METHOD OF DIAGNOSTIC TREES

In the experiment described in section 2, we had the luxury of performing a controlled comparison of the language of conversational speech with the written text of the Wall Street Journal. In many situations, however, we may not have the ability to isolate the contribution to the error rate of a single factor. In order to address the issue of understanding the contributions and interactions of multiple sources of error in an experiment, we have developed a general technique for identifying the attributes which distinguish correctly-recognized word occurrences from recognition errors. We quantify numerous characteristics both of the model and of the test utterances as a first step toward highlighting specific weaknesses in a given experiment.

For each unique word in the vocabulary we compile a set of measurements related to our ability to model that word effectively, independent of the context of neighboring words as well as of the channel. In addition, each utterance in the test set is characterized by a set of attributes related to the physical environment of the waveform or to the sequence of words it represents. These features can be thought of as context-dependent parameterizations of the test set. Each word in the test set is then characterized by the union of the context-independent and the context-dependent attributes.

In addition to being assigned a set of descriptive features, each word in the test set is assigned a class label, indicating whether or not the word occurrence was correctly recognized. The context-independent and context-dependent measurements for the word together with the class label are compiled into an MxN matrix, where M is the number of words in the test transcriptions and N-1 is the number of measurements compiled for each word.

Having assembled this matrix, we build a decision tree for the recognition experiment as a method of separating word occurrences which were correctly recognized from those which constitute recognition errors. The splits associated with the tree provide understanding of the major contributors to error in the experiment and serve as a guide for designing further recognition experiments. The data are partitioned along coordinate axes, rendering the tree particularly easy to interpret.

3.1. Feature Selection for Diagnostics

The specific features we model in order to build our diagnostic trees are described in this section. For diagnostic purposes, the statistics are compiled using the true transcription of each utterance. Using features compiled from recognition output as a precursor to predicting where errors have occurred in recognition will be considered in section 3.2.

Word-dependent (context-independent) features include the number of phonemes in the word, the number of times the word occurred in language model training as well as in acoustic training, and the minimum and average of the number of times the triphones which comprise the word occurred during acoustic training. In addition, we identify each word as being either "out of vocabulary," "non-speech," or "other."

The most informative of the context-dependent features has proved to be a word-level acoustic score, which is calculated by aligning a transcription to the corresponding acoustic models and averaging the log likelihoods of those frames aligning to any state within a given word to form the score for that word. Other contextual features include the likelihood of the sentence according to the language model normalized by the number of words in the sentence, an estimate of the signal-to-noise ratio, and the speaking rate, measured in terms of vowels per second as well as words per second. Phonemes per second was also considered as a measure of speaking rate but was not as good in predicting errors as the other measures. Finally, we include a smoothed score reflecting the number of times each word in its particular left and right context in the test set occurred in language model training. This feature, chosen to resemble the statistics used by the decoder, is given by $(\alpha L2+L3)(\alpha R2+R3)$ where $L2$ ($R2$) is the frequency of observing the word and its left (right) neighbor, $L3$ ($R3$) is the frequency of observing the word with its two left (right) neighbors in language model training and α is a back-off constant. We have used $\alpha = 0.1$ in our experiments.

As an example of a diagnostic tree we show in figure 1 the resulting analysis of our best performance on the Switchboard corpus to date [2]. The numbers at each leaf of the tree are the fraction of correctly-classified words mapping to that leaf. In this tree the word-level acoustic score is the main determiner of performance. The leaf with the best recognition accuracy (91%) represents those words which are long (4 or more phonemes), are spoken at a rate of less than 5.2 words per second, have a trigram occurrence score of more than 5.5, and have a word-level acoustic score greater than -4.33. Conversely, the words which have the lowest chance of being correctly recognized correspond to word-level acoustic scores less than -4.33, a trigram coverage score of less than 55.7, and are part of a sentence which has a language model likelihood of more than -2.61 but were themselves observed fewer than three times in acoustic modeling.

On occasion we have found it useful to include additional features in the analysis. This can easily be accomplished by adding a column to the existing matrix corresponding to the new attribute. Two features which proved to be informative from a diagnostic standpoint, as well as important in determining whether or not a word would constitute a recognition error were the identity of the speaker and a binary-valued variable indicating whether or not the adjacent word was correctly recognized.

Conversely, one may want to see the effect of excluding a feature from the input to the decision tree growing algorithm since a chosen split masks all others, even close contenders. For example, when we excluded the word-level acoustic score from consideration in building the tree of figure 1, speaking rate appeared as the first split in the resulting tree. This suggests that the underlying cause of the poor acoustic scores is rapid speech and the poor articulation which generally accompanies it.

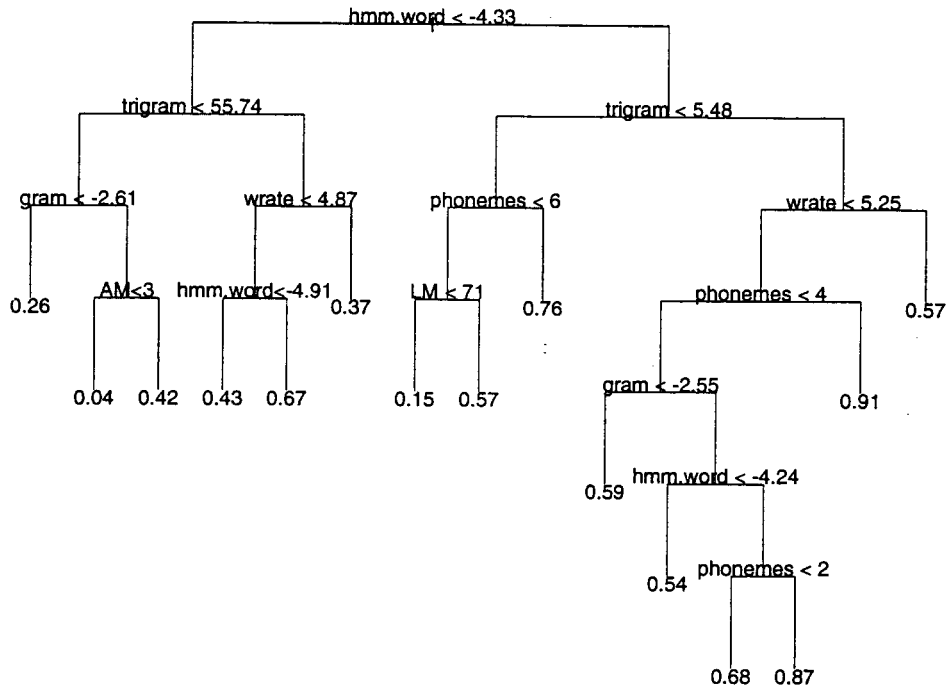

Figure 1: Tree resulting from decoding the males in the CAIP development test set with 66 hours of acoustic training. The true transcription of each utterance was used to discover the features. **hmm.word** is the word-level acoustic score. **trigram** is the backed-off measure of local coverage in language model training. **gram** is the sentence perplexity. **wrate** is a global estimate of speaking rate in words/sec. **phonemes** is the number of phonemes in each word. **LM** is the number of occurrences of the word in language model training. **AM** is the number of whole-word occurrences in acoustic training.

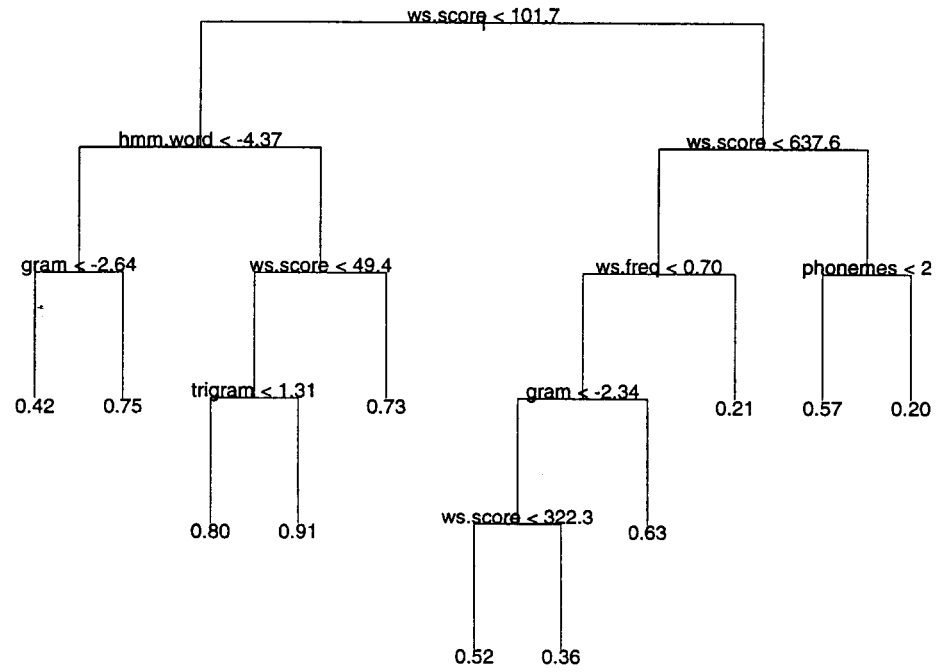

Figure 2: Tree corresponding to decoding the males in the CAIP development test set with 66 hours of acoustic training, built for assessing confidence in the recognition. The recognition hypotheses were used to discover all features. **ws.score** is the weighted wordspotting score. **ws.freq** is the unweighted wordspotting score. **hmm.word** is the word-level acoustic score. **gram** is the sentence perplexity. **trigram** is the backed-off measure of local coverage in language model training. **phonemes** is the number of phonemes in each word.

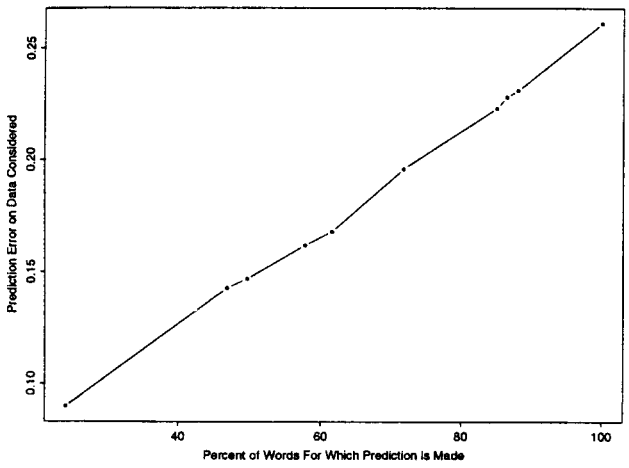

Figure 3: Errors in predicting whether a word is correct or incorrect as a function of the amount of data considered, considering nodes in the tree in order according to their purity.

3.2. Confidence Measures Through Diagnostics

In diagnosing the major sources of error in a particular experiment, we use the true transcription to derive features upon which to base our diagnostic tree. In addition to providing understanding, however, we would like to use the trees for predicting errors or assessing our confidence in the recognition output. In order to do so, we must be able to derive a meaningful set of features from the recognition rather than from truth.

The features described in section 3.1 are attributes which may be estimated using recognition rather than truth. For each unique word in our recognition output we measure the number of phonemes, the amount of language model training, the minimum and average triphone coverage as well as the number of whole-word occurrences in acoustic training, and whether or not the recognized item is non-speech. Aligning the recognition hypothesis to the acoustic models generates a word-level acoustic score for each word occurrence. By evaluating the language model we generate a sentence-level grammar score, and by the smoothed counting procedure outlined above we obtain local language model scores. We count the number of vowels and words in the hypothesis to derive estimates of speaking rate. Our SNR estimate does not rely on word identities.

In using a diagnostic tree for quantifying confidence in the recognition output, we have found two additional features to be helpful. These were not included in building trees for diagnostic purposes, as they provide little insight into the underlying causes of the errors. The additional features are wordspotting scores, calculated from the unweighted and weighted frequencies of each word in the recognition output in an N-best list. In our experiments we let N=100 and take the weights to be the decoder score of each hypothesis.

In figure 2 we show a diagnostic tree generated for the purpose of assessing confidence in the recognition. All features were derived from the recognition hypotheses. The new wordspotting features demonstrate much predictive power about the correctness of the hypotheses.

A diagnostic tree may be viewed as a density function which generates the probability of a word being correctly recognized given the recognition-dependent observations. Each leaf in the tree assigns to all words mapping to it a probability of being correct which is equal to the relative frequency of correctly-recognized words mapping to that leaf. Therefore, for each word hypothesized in recognition, we can assign an *a posteriori* probability of that word being correct. The individual scores indicate a confidence in our hypotheses beyond that possible through recognition scores alone.

Furthermore, we can form estimates of the incorrect words in a recognition hypothesis based on the probability of being correct given by the leaves of the diagnostic tree. Shown in figure 3 is a plot of the error incurred on the training data in predicting from the tree in figure 2 whether a word will be correctly or incorrectly recognized as a function of the number of predictions made. Initially predictions occur only at the leaf with the highest purity; subsequently lowering the threshold on node purity required before a prediction will be made increases the percentage of the test set for which a prediction is formed at the cost of a higher prediction error rate. The prediction error on the training data is 26.1% when all of the data is considered. When prediction is performed on all of the words in an independent test set of the features the error is 26.8%. Based only on the weighted wordspotting score the error on the training data is 30.2%. However, in the latter case, although setting thresholds on the score does provide a means of selecting subsets of the data on which to base predictions, direct use of classification error as a means of selecting the subsets for performance prediction is not possible.

4. CONCLUSION

In this paper we have reported the results of a diagnostic experiment which compared the recognition performance achieved on read Switchboard transcriptions with that achieved on WSJ sentences. We have observed a large increase in error rate on the Switchboard sentences, attributable chiefly to the shorter average word length on that corpus.

The remainder of the paper presented the technique of diagnostic trees for identifying the chief sources of error in a recognition experiment. The technique is provides understanding about the causes of error in a given experiment and therefore direction in designing the next experiment. Furthermore, it enables a measure of confidence in the recognition hypotheses.

We have envisioned several ways to incorporate the technique into our recognition system. As mentioned in the previous section, the diagnostic tree may be viewed as a density function, generating an estimate of the probability of being correct for each hypothesized word in an N-best list. The product of these likelihoods normalized by the number of words in the hypothesis could then be used to re-order the list. Similarly, diagnostic trees could prove useful in selective speaker adaptation, where the adaptation is based only on regions of hypothesized text in which we have high confidence. Another application of the technique lies in language modeling, where we might want to retreat from a trigram model when we are unsure of the output in hopes of a quick recovery from likely errors. Exploring these applications is the focus of our current research.

REFERENCES

[1] Schwartz, R. Personal communication. 1994.

[2] Jeanrenaud, P. et al. ICASSP 95 Proceedings.