BLAME ASSIGNMENT FOR ERRORS MADE BY LARGE VOCABULARY SPEECH RECOGNIZERS

Lin Chase

The Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, Pennsylvania 15213 USA
chase@cs.cmu.edu

ABSTRACT

This paper describes an approach to identifying the reasons that speech recognition errors occur. The algorithm presented requires an accurate word transcript of the utterances being analyzed. It places errors into one of the categories: 1) due to out-of-vocabulary (OOV) word spoken, 2) search error, 3) homophone substitution, 4) language model overwhelming correct acoustics, 5) transcript/pronunciation problems, 6) confused acoustic models, or 7) miscellaneous/not possible to categorize. Some categorizations of errors can supply training data to automatic corrective training methods that refine acoustic models. Other errors supply language model and lexicon designers with examples that identify potential improvements. The algorithm is described and results on the combined evaluation test sets from 1992-1995 of the North American Business (NAB) [1] [2] [3] corpus using the Sphinx-II recognizer [4] are presented.

1. Introduction

The main goal of this work, inspired in part by [5], is to provide new feedback mechanisms to speech recognizers by automatically identifying the sources of errors produced by the system. These errors are categorized in terms of components of the recognition system so that adjustments to the system, either automatic or human-supervised, can be made. The error blame assignment techniques discussed here are being developed in conjunction with a new type of automatic confidence annotation technique. The new confidence annotators don't simply identify errors as being present or not, but rather classify them (probabilistically) as being in one of several classes of error [6]. Confidence annotation for the purposes of error blame assignment theoretically will be much less expensive in the long run than the technique described here, as it does not require access to an accurate word transcript. The goal in the long run is to have such high quality multi-class confidence annotation that knowledge of the transcript becomes unnecessary for driving automatic error feedback. In the interim, however, the approach described here is useful.

While the main goal is as described above, another more subtle achievement is accomplished along the way. If we have a reliable error blame assignment technique that can tell us how many errors we could recover by working on a specific system component, then we are no longer tied to the simple word error rate (WER) measure when evaluating some particular design change in the system. For instance, if we know that only 15% of our errors on some test were caused by problems with the acoustic models, then we could be satisfied with a change that recovers most of these errors, even though our total reduction in WER would not be large. Thus a good blame assignment algorithm can give us the ability to accurately evaluate marginal improvements in system design.

2. Data Sources Used

The computations used by the blame assignment algorithm are based on several important sources of information. Some of these come directly from the internal workings of the speech recognizer. Others are additional features computed in parallel. These include:

- The contents of an N-best list, including complete word and phone segmentation and score information for each of its 150 elements, the first of which is the system's best HYPothesis,
- Language model score and source information from the same best-scoring HYPothesis,
- Word and phone segmentations and acoustic scores from the Viterbi alignment of the utterance REFerence transcript,
- Language model score and source information, calculated from applying the recognition language model to the REFerence transcript,
- Characteristics about the words HYPothesized, especially their dictionary pronunciations,
- The results of a parallel "phone-only" decoding, in which the recognizer is not constrained to either phone sequences from the dictionary or word sequences in the language model in its recognition of phone sequences,
- Three distance metrics between basephones, including a simple match count at the frame level, a phonologically-based similarity measure (H_{WC})[6], and an empirically derived confusion-based distance measure,

3. Defining Error Regions

A detailed description of a technique in which error regions are identified and analyzed can be found in [6]. What follows is a summary of this technique.

An error region is a contiguous set of frames of acoustic data. It starts at the beginning of a word and ends at the end of a (possibly distinct) word. An error region is identified by comparing the REFerence of an utterance against HYPothesized recognition output. In its strictest configuration, the process that identifies error regions requires that all word segmentations match those in the reference exactly. That is, for each word segment in the reference, the hypothesis must contain a segment with the identical word (including pronunciation variant), start frame, end frame, and thus acoustic score. It is possible that an error region may contain only one word. In the most extreme case, an error region might contain all the words in the utterance.

This definition of *error region* also included two additional factors:

1. A *frame tolerance* can be specified that allows some slack in the comparisons between boundary locations in the reference and hypothesis sequences.
2. The criterion for locating the end of an error region can be changed to reflect the effects of the language model being used by the recognizer. The *window size* parameter can be set to a value of 0, 1, or 2. These values correspond respectively to ignoring language model effects, assuming a bigram language model, and assuming a trigram language model.[1] The parameter is interpreted as the number of successor words that are included in the error region after the hypothesis returns to matching the reference.

Within each identified error region, the following information is collected for later evaluation:

1. *Acoustics:*
   - the total acoustic score of the hypothesized words in the error region,
   - the total acoustic score of the reference words in the error region,
   - the difference between the two, noting whether the reference or hypothesized acoustic score is better (either HYP/AC or REF/AC).
2. *Language model:*
   - the total language model score of all of the words in the hypothesis that fall into the error region,
   - the total language model score of all of the reference words in the error region,
   - the difference between these two, noting which is larger (either HYP/LM or REF/LM).
3. *Totals:*
   - the total combined acoustic and language model score for the hypothesis portion of the error region,
   - the total combined acoustic and language model score for the reference segments in the error region,
   - the difference between these two totals, noting which is larger (either HYP/TOT or REF/TOT).

Using these values is possible to classify error regions as falling into one cell of a 2-by-3 matrix, as shown in Table 1. The column of the table in which the error region will be placed is determined by whether the REF or HYP had the better overall score (determined by whether REF/TOT or HYP/TOT was chosen). The row is determined by which of acoustic, language, or both acoustic+language caused the HYPothesis to be chosen (determined by REF/LM vs. HYP/LM and REF/AC vs. HYP/AC).

For an example of how this works, consider the error region displayed in Figure 1. In this case the acoustic models preferred the reference sequence. On the other hand, the hypothesized word sequence was preferred by the language model. This language model preference was strong enough to overcome the otherwise correct acoustics and create an error. This error region is categorized in the "HYP language model dominates REF acoustics" cell of Table 1. The total score of the hypothesis portion of the error region was better by 48 points. (See [6] for a graphic version of the placement of error regions w.r.t. the cells of Table 1.)

[1] This assumes forward N-gram models. The extension would be in the opposite direction for backward N-grams.

	REF total better	HYP total better
AC bigger	REF acoustics dominate HYP language model	HYP acoustics dominate REF language model
LM bigger	REF language model dominates HYP acoustics	HYP language model dominates REF acoustics
LM bigger	REF acoustics and language model both better than HYP	HYP acoustics and language model both better than REF

Table 1: Error regions can be categorized as belonging to one of these six categories, based on the HYP and REF acoustic and language model scores.

We can gain more insight if we look at what caused the errors using the blame assignment algorithm described in the next section.

4. Classifying Error Regions

Figure 2 describes a control flow sequence that we pass through in applying the blame assignment algorithm. What follows is a stepwise description of each element of the chart. First we analyze the utterance as described above to produce a set of error regions. We then proceed with the following tests. As soon as a category for an error region has been found we go on to the next region.

1. *OOV errors*: Is there an out-of-vocabulary (OOV) word in the reference portion of the error region? If so, then we categorize the error region and all of the word errors it contains as belonging to class *OOV*.
2. *Search errors*: Did the reference portion of the error region receive a better score overall than the hypothesis portion? Then if a full search had been used instead of the suboptimal methods we rely on, this answer would have been found. Thus we categorize this error as belonging to class *search*.
3. *Homophone substitutions*: Did the reference and hypothesis portions of the error region contain the exact same phone sequence and acoustics score, yet the hypothesis was chosen over the reference? Then the hypothesis portion of the error region contains a homophone word or word sequence to that found in the reference portion, and the language model chose it incorrectly. Thus we categorize this error as belonging to class *homophone substitution*.
4. *Language model overwhelms*: Did the reference acoustics get a better score than the hypothesized acoustics, yet the hypothesis was chosen anyway? Then the acoustic models were capable of making the right decision, but the weighted language model probability was too large to let this happen. In this case we can distinguish between two subcases by asking:
   (a) *Language model overwhelm, LW adjustment possible*: Is there a possible pair of language weight/insertion

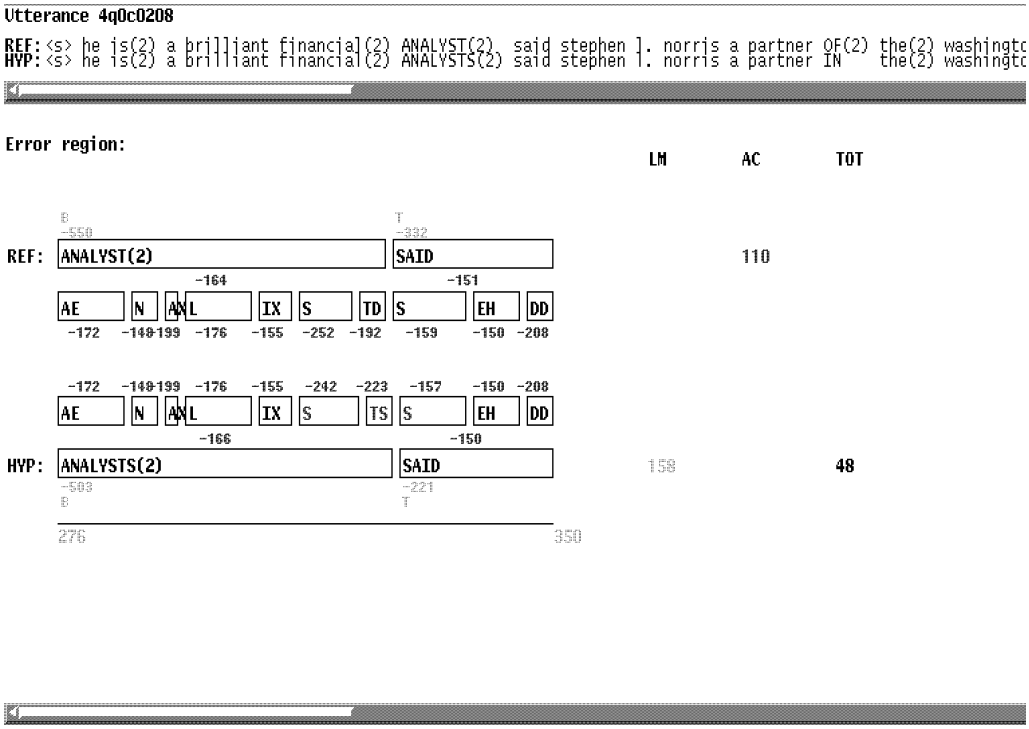

Figure 1: An error region caused by the hypothesis language model score overwhelming the otherwise correct reference acoustics.

penalty (LW/IP) for which the weighted language model probability becomes small enough to let the acoustics make the correct choice? [2] If so, then we categorize this error as belonging to class *Language model overwhelm: adjustment possible*.

(b) *Language model overwhelm*: If changing the weights on the language model probability cannot let the correct acoustic decision emerge, then we categorize the error as belonging to class *Language model overwhelm.*

5. *Both acoustic and language model overwhelm*: Did the hypothesis portion of the error region score better both according to the acoustic models and the language model? If so, then we categorize the error as being due to *Both acoustic and language model overwhelm.*

6. *Multiple types of acoustic problems*: At this stage we know that the language model score of the hypothesis was not better than that of the reference. We also know that that the hypothesis acoustic score was better than the reference acoustic score. When we reach this point we apply a variety of acoustic tests to try to determine whether:

- the reference acoustic score is somehow not accurate, or
- the acoustic models preferred the hypothesis incorrectly.

Using the tests described next, it is not always possible to tell the difference between these cases. Thus at this point we have a third option, as well: the default categorization "Miscellaneous".

The acoustics tests proceed as follows:

(a) *Pronunciation Missing/Transcript Error*: Does the REFerence describe accurately what was really said? Maybe the transcript is wrong, or maybe the pronunciation of the word used is missing from the dictionary. To find out if we're in this class we need to look for gross differences between what we expected to see (based on the phone-only decoding and what's in the hypothesis and N-best list) and what we did see (the HYPothesis). We can also look for evidence that the aligner was straining to fit the REFerence transcript to the acoustic data. (Details of how this is computed can be found in [6].) If this set of tests do not pass, then we move to the next round of tests.

[2] This calculation was not available in the experiments reported here.

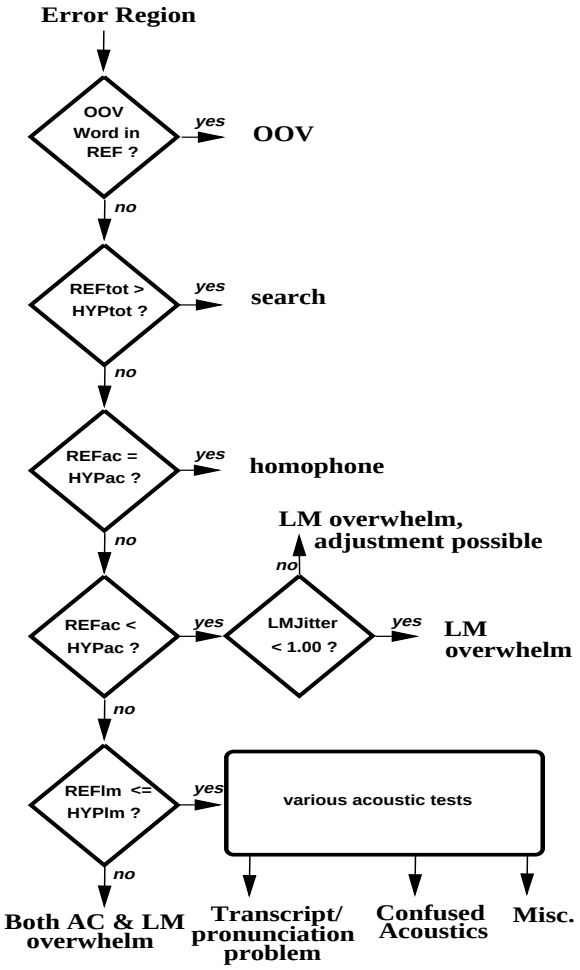

Figure 2: The sequence of tests applied by the blame assignment algorithm.

(b) *Confused Acoustics*: Did some phone model or models give too high a score to the wrong answer? Or perhaps, did the right answer's models not respond with a high enough score? To find out if we are in this case we need to test for evidence of acoustic confusion on the part of the acoustic models. To do this we test the earliest (leftmost) position in the error region in which the hypothesis and reference do not match. (Details can be found in [6].)

(c) If neither of the above work out, then we categorize the error as belonging to class *Miscellaneous*.

Table 2 shows the distribution of categorizations for the NAB development data, a total of 1523 utterances.

Category	%Regions
OOV	33.7%
search	9.9%
homophone substitution	2.4%
LM overwhelm	12.3%
Transcript/pronunciation	0.3%
Acoustic Confusion	10.2%
AC+LM Overwhelm	14.7%
Miscellaneous	21.5%

Table 2: Categorization of the error regions and errors found in the NAB test data.

5. Summary

We presented an approach to categorizing the errors that occur during speech recognition. We discussed how regions of error are identified, and how the acoustic and language model scores used in the recognizer can be compared with similar scores generated for reference transcripts under forced alignment. A general six-way categorization of error regions based on this approach was discussed. A blame assignment algorithm which refines this approach further was then presented. This algorithm places errors into one of the following categories: 1) due to out-of-vocabulary (OOV) word spoken, 2) search error, 3) homophone substitution, 4) language model overwhelming correct acoustics, 5) transcript/pronunciation problems, 6) confused acoustic models, or 7) miscellaneous/not possible to categorize.

The possible uses of the various categorized errors include both the feeding back of errors to acoustic corrective training algorithms and the possibility of further human analysis for appropriate adjustments to language models, transcripts, and dictionaries. Figures for the distribution of error types in the NAB development test set were presented.

References

1. Paul, D. and Baker, J. *The Design for the Wall Street Journal-based CSR Corpus*. in: **DARPA Speech and Language Workshop**. Morgan Kaufmann Publishers, San Mateo, CA, 1992.
2. Pallett, D., Fiscus, J., Fisher, W., Garofolo, J., Lund, B., and Przybocki, M. *1993 Benchmark Tests for the ARPA Spoken Language Program*. in: **ARPA Speech and Natural Language Workshop**. 1994, pp. 15–40.
3. Pallett, D., Fiscus, J., Fisher, W., Garofolo, J., Lund, B., Martin, A., and Przybocki, M. *1994 Benchmark Tests for the ARPA Spoken Language Program*. in: **ARPA Spoken Language Systems Technology Workshop**. 1995, pp. 5–38.
4. Hwang, M.-Y. *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*. School of Computer Science, Carnegie-Mellon University, December 1993.
5. Eide, E., Gish, H., Jeanrenaud, P., and Mielke, A. *Understanding and Improving Speech Recognition performance Through the Use of Diagnostic Tools*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1995, pp. 221–224.
6. Chase, L. L. *Error-Responsive Feedback Mechanisms for Speech Recognizers*. Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh PA USA, also available as Tech Report CMU-RI-TR-97-18, April 1997.