BETTER ALIGNMENT PROCEDURES FOR SPEECH RECOGNITION EVALUATION

William M. Fisher and Jonathan G. Fiscus

National Institute of Standards and Technology (NIST)
Gaithersburg, MD 20899

ABSTRACT

In evaluating speech recognition, an alignment of reference (REF) symbols with hypothesized (HYP) symbols is the basis of other measures. It is therefore important to do this keystone step well. We report here on recent advances made at NIST on algorithms for alignment, empirically justifying "phonological" alignment, which minimizes differences in phonological features, and briefly describing a new technique for identifying "splits" and "merges".

1. INTRODUCTION.

The most generally accepted method of scoring and evaluating speech recognition consists of comparing a transcription that is assumed to accurately represent what the speaker said, the reference ("REF") transcription, with a transcription representing the speech recognizer's hypothesis as to what the speaker said, the hypothesized ("HYP") transcription. An alignment between units in the two transcriptions is found, and then evaluative scores are tallied, based on the alignment. The scores are typically counts of insertions, deletions, and substitutions, and words are typically the units transcribed. Various helpful diagnostics can also be based on these alignments. Figure 1 below illustrates a typical alignment; asterisks stand for the null element, and the alignment indicates one substitution and one deletion error.

```
REF: | the | best | of | times |
HYP: | the | test | ** | times |
ERRS:       S      D
```

Fig. 1. A Typical Alignment for Speech Recognition Evaluation.

This alignment indicates one word substitution ("test" for "best") and one deletion ("of"). Note that the HYP word "test" could have been aligned with the REF word "of" without affecting the gross word error count; and at most one-to-one correspondences can be indicated. The major problems with this kind of alignment are: 1. what units to use; 2. whether to allow one unit to match several; and 3. how to find the right alignment. In this paper, we focus on the last of these concerns.

Doing alignments correctly is important because alignments are the bases of most if not all speech recognition evaluation and diagnostics. This paper reports further justification for the NIST phonological feature-based alignment software first described briefly at ICASSP 90 [1].

2. PREVIOUS WORK.

The problem of finding an alignment is solved by finding one that optimizes an objective function of the correspondences it indicates, commonly a weighted sum of insertions, substitutions, and deletions, and the usual approach to finding the solution is a dynamic programming algorithm [2]. The NIST software that is currently being used in the DARPA-supported speech recognition efforts minimizes the sum of word substitutions weighted by 4 and insertions and deletions weighted by 3, with no other attempt to align similar words. This weighting was chosen to make the algorithm prefer a substitution over the logically equivalent pair of a deletion and an insertion.

This classical approach sometimes indicates word substitutions that seem extremely unlikely, the sounds of the matched words having little in common. Attempts to make the word alignment more sensitive to the sounds of the words have been reported by Picone et al. [3,4] In this work, simple DP word alignment schemes such as the current NIST/DARPA one are termed "word-mediated" and their method is called "phone-mediated", because it first aligns strings of phones representing the utterances and then backtracks to align words. (By analogy, we should call the new method reported on here "feature-mediated" alignment, but we are used to calling it "phonological".)

Hunt [5] has shown that word-mediated alignment can slightly though significantly underestimate true error rate and can be somewhat biased.

3. THE PHONOLOGICAL ALIGNMENT ALGORITHM.

This way of aligning transcriptions is called "phonological" because it finds the alignment of words that minimizes "phonological distance", currently calculated as the minimum number of phonological features that must be changed to turn one transcription into the other. The notion that phonological feature representation provides a language-universal space in which utterances that sound similar are nearer is implicit in much of the modern linguistic literature. In the version tested here, the phonemic representations of words are taken from a dictionary distributed under license by Janet Baker of Dragon Systems, backed up by a crude letter-to-phoneme function, and feature representations of the phonemes are a highly modified set ultimately derived from one sent to us by Mike Cohen of SRI. A particularly good example of the differences that can be expected between phonological alignments and word-mediated ones is shown below in Figure 2; not all are so favorable.

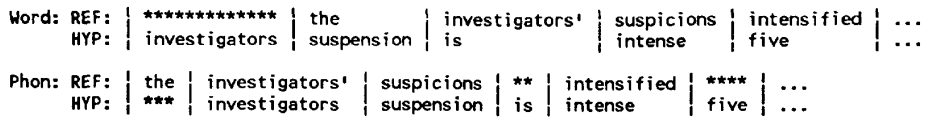

Fig. 2. Example of Word-mediated vs. Phonological Alignments.

4. SUBJECTIVE PREFERENCE TEST

An experiment was done to test the hypothesis that expert speech scientists generally prefer phonological alignments over simple word-mediated alignments.

4.1 Experimental Design and Procedure

The test used a forced-choice, matched-pairs, double-blind experimental design, in which for each item the two different alignments of the same sentence were presented to each of the judges for preference rating. The two methods were (W) the word-mediated alignment; and (P) the phonological alignment. In the actual test materials, the order of presentation of the two methods was randomized and unknown to the subjects. Subjects were just asked to indicate which alignment they preferred. Test materials were generated, randomized, and analyzed by computer programs; the experimenter didn't know which answers were the "right" ones. The 25 sentences were selected at random from those having different W and P alignments in the DARPA February '92 Wall Street Journal (WSJ) corpus speech recognition test and training materials. The order of presentation of the test items was also randomized.

4.2 Subjects

The 13 subjects ("judges") were experts in the field of speech recognition: Janet Baker, Jared Bernstein, George Doddington, Melvyn Hunt, Lauri Lamel, Chin-Hui Lee, John Makhoul, Mari Ostendorf, Dave Pallett, Mike Phillips, Patti Price, Bob Weide, and Victor Zue.

4.3 Results.

Table 1 below shows the W/P preferences for each judge, summing over utterances, along with an estimate of the statistical significance of their preferences (judges are ordered differently than in the list above).

The probability of the null hypothesis was computed using the 1-tailed (directional) sign test for matched pairs [6], with the normal approximation to the binomial distribution and calculating z with the continuity correction.

4.4 Summary.

The participating judges unanimously prefer the results of the new phonological method over the word-mediated method. For 12 of the 13 judges, this is statistically significant at the $p < .001$ level; for judge #8, the significance level is $p < .03$.

judge	W	P	z	p
1	4	21	3.20	<.000687
2	4	21	3.20	<.000687
3	3	22	3.60	<.000159
4	3	22	3.60	<.000159
5	3	22	3.60	<.000159
6	2	23	4.00	<.000032
7	0	25	4.80	<.000003
8	7	18	2.00	<.022750
9	1	24	4.40	<.000032
10	2	23	4.00	<.000032
11	1	24	4.40	<.000032
12	1	24	4.40	<.000032
13	2	23	4.00	<.000032
Σ	33	292		

Table 1. Judges' Alignment Preferences.

5. ENTROPY TEST

In response to a criticism that the phonological method would "add noise to the scoring", we computed the entropy of the sets of word confusion pairs (including identical word matches) indicated by the word-mediated and phonological alignment methods applied to all the sets of data from the February 1992 DARPA Wall Street Journal speech recognition tests. The results are presented below.

As can be seen, in every case the entropy either decreased or stayed the same, with a mean decrease of 0.65%. The entropy of the confusion sets is less with the phonological alignments than with the word-mediated ones ($p < .005$). The statistical significance of this result was assessed by a 1-tailed t-test for correlated samples, $t = 5.18$, $df = 21$. This indicates that the phonological method generally reduces noise in the alignments, which should improve the quality of diagnostics based on them.

	ENTROPIES			
set	Word	Phon	diff	%diff
1	7.94	7.94	0.00	0.0
2	8.01	8.01	0.00	0.0
3	7.94	7.94	0.00	0.0
4	8.69	8.68	0.01	0.12
5	8.70	8.68	0.02	0.23
6	8.12	8.12	0.00	0.00
7	8.82	8.79	0.03	0.34
8	8.25	8.25	0.00	0.00
9	9.11	9.07	0.04	0.44
10	8.92	8.88	0.04	0.45
11	8.97	8.95	0.02	0.22
12	9.24	9.21	0.03	0.32
13	9.15	9.09	0.06	0.66
14	9.14	9.05	0.09	0.98
15	9.44	9.35	0.09	0.95
16	9.55	9.45	0.10	1.05
17	9.89	9.77	0.12	1.21
18	9.33	9.24	0.09	0.96
19	10.05	9.93	0.12	1.19
20	9.42	9.30	0.12	1.27
21	9.80	9.67	0.13	1.33
22	9.94	9.77	0.17	1.71
Avg	9.02	8.96	0.06	0.65

Table 2. Confusion Word-pair Entropy Under Word and Phonological Alignments.

6. COMPARISON WITH TIME-REGISTERED ALIGNMENT

Another approach to evaluating two competing alternatives is to compare each to an assumed ideal standard. The ideal that we worked with uses beginning and ending time registration marks for each word in the REF and HYP strings, minimizing an objective function of the time mis-match (as in Hunt 1990, op cit.). The particular word-to-word distance function that we used is the absolute value of the difference in beginning times plus the absolute value of the difference in ending times. Time-marked HYP files were kindly supplied to us by Hy Murveit of SRI, and we time-marked the REF files by hand.

6.1 Experimental Design.

The first 15 utterances that SRI sent us were used in this experiment. Both word-mediated and phonological alignments were created and compared to the time-registered ("T") alignments. In each comparison, we counted the number of word slots that did not match the T alignment ("Nwt" for the word alignment, "Npt" for the phonological), using the same dynamic programming algorithm that is used in the simple word alignment function. The statistic that we used to estimate the significance of the results was the signed difference, (Npt-Nwt).

6.2 Results.

Table 3 shows the results of the comparisons:

#	utt id	Nwt	Npt	diff
1	060o2001	0	0	0
2	060o2002	0	2	2
3	060o2003	5	5	0
4	060o2004	0	0	0
5	060o2005	0	0	0
6	060o2006	0	0	0
7	060o2007	0	0	0
8	060o2008	0	0	0
9	060o2009	2	2	0
10	060o200a	0	0	0
11	060o200b	0	0	0
12	060o200c	5	8	3
13	060o200d	0	0	0
14	060o200e	0	0	0
15	060o200f	3	5	2
	Σ	15	22	7

Table 3. Comparision of Word, Phonological, and Time-Mark Alignments. Nwt is the number of word-slot errors comparing the word alignment to the time-mark alignment; Npt is the same measure comparing the phonological alignment to the time-mark one; diff is (Npt - Nwt).

6.3 Summary.

The results indicate that the phonological alignments are significantly closer to these ideal time-mark alignments than are the word alignments ($p < .05$). The statistical significance of this result was assessed by a 1-tailed t-test for correlated samples, $t = 1.82$, $df=14$. (As expected from Hunt's work (op cit.), the time-marked alignments indicate slightly lower word accuracy.)

6. SUMMARY.

We have presented three justifications that argue for use of the new phonological alignments over the current word-mediated ones:

1. Experienced speech scientists generally prefer them;
2. They reduce the entropy of the indicated matches;
3. They agree more often with ideal alignments based on time registration.

7. FURTHER WORK

All of the alignments compared here constrain the indicated matches to be at most one-to-one. We are currently exploring algorithms for handling "splits" and "merges", which allow one-to-many matches, using the criterion that, e.g., a split is to be preferred to the equivalent combination of a substitution and an insertion if the phonological distance indicated is less. We hope to report results at a later conference.

8. ACKNOWLEDGEMENTS

We thank Paul Michaelis of ATT and Gerry Birdwell of TI for trying to help us avoid some of the pitfalls of human-factors experimental design.

REFERENCES

[1] D. S. Pallett, W.M.Fisher, and J.G. Fiscus, "Tools for the Analysis of Benchmark Speech Recognition Tests", *Proc. ICASSP*,1990, pp. 97-100.

[2] D. Sankoff and J.B. Kruskal (eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesly, Inc., Reading, 1983 (ISBN 0-201-07809-0).

[3] J. Picone, G.R. Doddington, and D.S. Pallett, "Phone-Mediated Word Alignment for Speech Recognition Evaluation", *IEEE Trans. ASSP*, Vol. 38, No. 3, March 1990, pp. 559-562.

[4] J. Picone, K.M. Goudie-Marshall, G.R. Doddington, and W. Fisher, "Automatic Text Alignment for Speech System Evaluation", *IEEE Trans. ASSP*, Vol. ASSP-34, NO. 4, August 1986, pp. 780-784.

[5] Melvyn J. Hunt, "Figures of Merit For Assessing Connected-word Recognisers", *Speech Communication* Vol. 9 (1990), pp. 329-336.

[6] Robert L. Winkler and William L. Hays, *Statistics: Probability, Inference, and Decision*, 2nd ed., Holt, Rinehart and Winston, New York, 1975, p. 855.