Published in *IEEE International Conference on Speech and Language Processing (ICSLP)*, Pekin, pages II:915–918, oct 2000

Combining Multiple Speech Recognizers using Voting and Language Model Information

Holger Schwenk and Jean-Luc Gauvain

{schwenk,gauvain}@limsi.fr
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE

ABSTRACT

In 1997, NIST introduced a voting scheme called ROVER for combining word scripts produced by different speech recognizers. This approach has achieved a relative word error reduction of up to 20% when used to combine the systems' outputs from the 1998 and 1999 Broadcast News evaluations. Recently, there has been increasing interest in using this technique. This paper provides an analysis of several modifications of the original algorithm. Topics addressed are the order of combination, normalization/filtering of the systems' outputs prior to combining them, treatment of ties during voting and the incorporation of language model information. The modified ROVER achieves an additional 5% relative word error reduction on the 1998 and 1999 Broadcast News evaluation test sets. Links with recent theoretical work on alternative error measures are also discussed.

1. INTRODUCTION

The National Institute of Standards and Technology (NIST) has a long tradition in organizing evaluations of LVCSR systems. In 1997, NIST presented an approach, named ROVER (*Recognizer output voting error reduction*), to combine the transcribed outputs of several recognizers in order to produce new (improved) transcriptions [1]. ROVER was first used to combine the results submitted by all participants in the LVCSR 1997 Hub 5-E evaluation: the word error rate was reduced from 44.9% (for the best single system) to 39.4%. This approach has since gained increasing interest with five of the nine participants in the 1998 Broadcast News evaluation submitting a speech recognizer that itself is a combination of several different recognizers. Despite this, NIST was still able to reduce the error rate from 13.5% to 10.6% by performing ROVER on the nine participating systems [5].

Recently, links of the ROVER algorithm with theoretical work on n-best-list or lattice-based word error minimization [3, 6] and task-dependent error measures [2] have been established. To the best of our knowledge, however, there has been no implementation and large scale evaluation of a modified ROVER algorithm. We believe that there are many open questions, for instance, how important is the order of combination ? how many systems should be combined ? is it advantageous to preprocess or normalize the systems' outputs prior to combination ? and should we incorporate language model information into the combination process ? In this paper we report results that attempt to clarify these questions. In the next section we summarize the ROVER algorithm. Section 3 then describes some extensions of the core algorithm and we present results on the Broadcast News 1998 and 1999 evaluation set. The paper concludes with a discussion of future research issues.

2. ROVER

ROVER was developed by J. Fiscus of NIST [1]. It seeks to reduce word error rates for automatic speech recognition by exploiting differences in the nature of the errors made by multiple speech recognizers. Rover proceeds in two stages: first the outputs of several speech recognizers are aligned and a single word transcription network (WTN) is built. The second stage consists of selecting the best scoring word (with the highest number of votes) at each node. The decision can also incorporate word confidence scores if these are available for all systems.

It is quite difficult to optimally align more than two word sequences and an iterative procedure is used. First, two sequences are aligned, creating a combined word transcription network. This WTN is aligned with the third word sequence giving a new combined word transcription network, that itself is aligned with the fourth word sequence and so on. The use of no-cost word transitions ("@"-arcs) allows insertions and deletions to be handled (see [1] for more details). Note that decisions are made **separately** at each node based on local information, i.e. the number of occurrences and/or the confidence score of each alternative arc. This means in particular that no information about the word context is used and as a result the combined output may have a very high perplexity. This is in contrast to the usual approach to speech recognition where language model (LM) information tends to reduce the perplexity of the hypotheses.

3. ANALYSIS AND EXTENSIONS

Table 1 gives the results of all the participants in the 1998 DARPA Broadcast News evaluation [5]. ROVER is the result of combining all the nine systems in *alphabetical* order. Recall also that four of the five best systems already used ROVER (ibm, cu-htk, dragon and bbn).

ibm	limsi	cu htk	dragon	bbn	philips rwth	sprach	sri	ogi fonix	**ROVER**
13.5	13.6	13.8	14.5	14.7	17.6	20.8	21.1	25.7	**10.6**

Table 1: Official word error rates in % for the 1998 Broadcast News evaluation set (after [5]).

Order of combination

It is known that the pairwise alignment procedure of ROVER is to some extent affected by the order of combination. Furthermore, ROVER is here used to combine outputs of continuous speech recognizers, that means a sequence of words without any sentence structure. For efficiency reasons, during the alignment process it is necessary to split one document into smaller parts (for Broadcast News, each document contains more than 14k words). This is done by searching for gaps larger than one second in the first word sequence. The document is then split at this point if there is a corresponding silence in all other word sequences. Obviously, the results depend on which word sequence is used first. Therefore, it can be advantageous to use the best single word recognizer as the first system, and more generally, to combine them in the order of decreasing recognition rate. We combined the outputs according to the performance of each speech recognizer on the 1997 Broadcast News evaluation set. Figure 1 (solid line) shows the word error rates when the recognizers are combined in error ranked order.

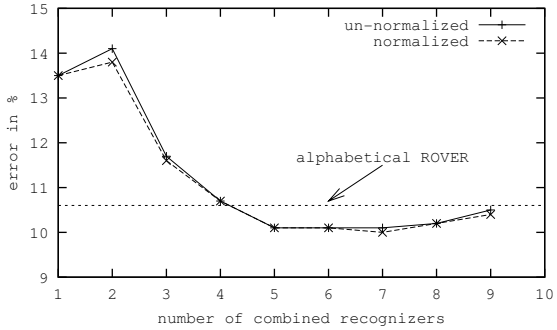

Figure 1: 1998 Broadcast News word error rates in function of the number of combined systems (individual error ranked order).

Although the combination of nine systems in ranked order instead of alphabetical order achieves only a very slight improvement in word error to 10.4%, a clear minimum of about 10.1% can be obtained when combining 5 to 8 systems. It appears that combining many systems, in particular those with higher error rates, is of no benefit and may actually increase the error rate of the combined system.

Normalization/filtering

The standard NIST scoring procedure applies a filtering/normalization of the recognizer's output prior to alignment with the reference transcription. This normalization includes mappings of alternative spellings to one common form (e.g. `afterall` $\rightarrow$ `after all`, `cannot` $\rightarrow$ `can not`, ...), and mappings of abbreviated forms to several variants (e.g. `CHILD'S` $\rightarrow$ `CHILD'S` or `CHILD IS` or `CHILD HAS`). We suggest applying this filtering **before** combining the systems with ROVER. The alignment of word sequences with variants, however, is not supported by ROVER. The dashed line in Figure 1 shows the resulting word error rates when the one-to-one filtering rules are applied. Surprisingly, there is only a slight decrease in the word error rate (10.1% to 10.0% when combining the outputs of seven recognizers). We conjecture that more improvements could be expected if all the filtering/normalization rules would be applied.

We have reimplemented the ROVER algorithm in order to add the extensions described in the following sections, in particular the incorporation of language model information. The program can combine the nine systems in 0.01xRT on a SGI UNIX workstation. Research on alignment procedures supporting variants or n-best lists as input is currently in progress.

Treatment of ties

When combining several systems it is quite frequent that after alignment some words appear equally often at a given node in the WTN and an arbitrary decision has to be taken (see Table 3).

# of combined recognizers	2	3	4	5	6	7	8	9
# of ties	4726	1560	1539	987	1075	884	923	846

Table 3: Number of ties for 1998 Broadcast News.

These ties could be broken using confidence scores of the individual systems, but unfortunately only three of nine participants of the 1998 Broadcast News evaluation provided them, so that this option wasn't possible. Also, the confidences scores provided by different recognizers may be difficult to compare.

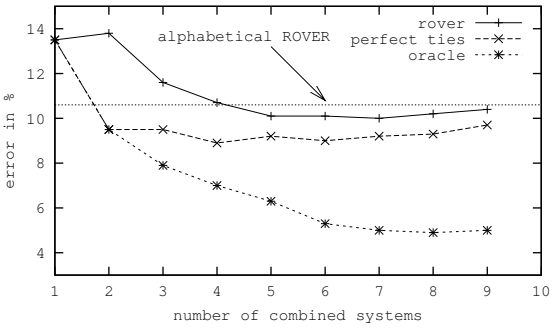

Figure 2: 1998 Broadcast News word error rates when deciding ties arbitrarily („rover"); making the best choice among the ties („perfect ties") and using the best fitting sequence in the whole aligned WTN („oracle").

Instead, we have determined the error rate that could be obtained if we correctly broke all the ties (see Figure 2 dashed line). In this case the word error rate would be about 9% which would be a significant improvement. In the next section, we present an approach how to break the ties using language model information. To determine the limit of this type of approach we show in Figure 2 (dotted line) the error rate that could be achieved if we always chose the correct word at each branch among all the alternatives (oracle-mode). It is surprising to see that the combined transcriptions form the nine speech recognizers contain the correct word over 95% of the time. These results are of course only of hypothetical value, but it seems nonetheless that there is some hope for further improvement of the combination approach.

Importance of language model information

One of the mysteries about the success of ROVER is that it seems to work well even though that no context and language model information is used. In fact, it could theoretically happen that the resulting word sequence has a higher perplexity than any of the individual word sequences. Therefore, we propose using

number of combined systems:	2	3	4	5	6	7	8	9
arbitrary ties:								
word error:	13.8%	11.6%	10.7%	10.1%	10.1%	10.0%	10.2%	10.4%
sentence error:	81.0%	76.3%	74.3%	73.0%	73.8%	73.4%	73.4%	74.6%
perplexity:	183.8	171.6	166.1	164.2	161.7	160.2	159.3	159.6
using LM to break ties:								
word error:	12.5%	11.1%	10.3%	10.1%	10.1%	10.0%	10.3%	10.5%
sentence error:	79.9%	75.4%	73.3%	72.6%	73.0%	72.9%	74.2%	74.7%
perplexity:	137.2	145.8	146.5	151.2	149.6	150.8	150.0	151.1

Table 2: 1998 Broadcast News test set word error rates and perplexity when using LM information instead of braking ties arbitrarily (see text for more details). NIST's ROVER achieves 10.6% word error and 73.7% sentence error.

LM information to achieve further improvements. This is done in the following way: first all the systems are aligned and the most likely word is selected at each branch of the word transition network. If several words are equally frequent, all are kept. Second we use the language model of LIMSI's Broadcast News system in order to select the word sequence among all alternatives that minimizes the perplexity.

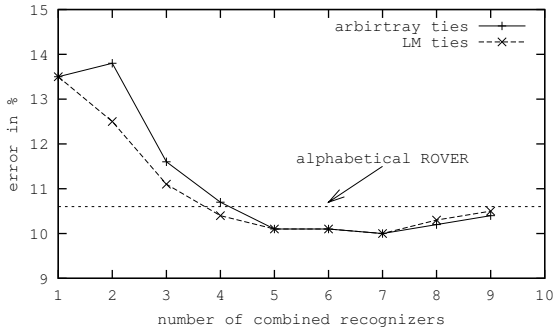

Figure 3: 1998 Broadcast News word error rates when using LM information instead of braking ties arbitrarily (see text for more details).

Table 2 gives the improvements of the word and sentence error rate as well as the perplexity when using a LM to break ties. An interesting result is obtained when combining just two systems: 8.1% relative word error reduction with respect to the best two individual systems (13.5 and 13.6% word error rate respectively). Note that in this case we always get ties when the two systems disagree which means that the LM is used for the whole decision process. Standard ROVER, e.g. breaking ties arbitrarily, does not work when combining just two systems (the word error increases to 13.8%). As can be seen in Figure 3, the use of LM information to break ties always gives better results than an arbitrary decision, but it seems to be particularly interesting when only few recognizers are combined: for instance a word error rate of 11.1% is achieved when combining the three best recognizers.

To the best of our knowledge, similar modifications of the reference ROVER algorithm have not been reported in the literature. There is however related work on hypotheses selection during decoding for a single speech recognizer [2, 3, 6]. In the standard approach to speech recognition, the goal is to find the sentence hypothesis that maximizes the posterior probability $P(W|A)$ of the word sequence W given the acoustic observation A. Usually speech recognizers are evaluated by measuring the word error so that there is a mismatch between the training and the evaluation criterion. Recently, algorithms for minimizing directly the word error have been proposed [2, 3, 6]. These approaches have been evaluated on the Switchboard corpus and achieved a small but consistent decrease in word error and an *increase* of the sentence error, in accordance with the new optimization criterion.[1] It is believed that word error minimization is most effective on tasks with relatively high error rates since a wrong sentence probably contains several wrong words.

We have not observed an increase in the sentence error, neither when using the original ROVER algorithm nor when incorporating LM information (see Table 2). In contrast to the above cited approaches to hypothesis selection in a single speech recognizer output, only limited information is available when applying ROVER: one single transcription with timing information for each speech recognizer. The only information that can be used is the number of occurrences of each word at a given node which was demonstrated to lead to suboptimal results when breaking ties arbitrarily. Our proposal to use a LM to break these ties combines a word error oriented criterion (local number of occurrences) with a sentence error criterion (minimum perplexity of the global word sequence). For comparison, we have also used the LM on the whole WTN, that means disregarding all information on the number of occurrences of each word. As expected, the results were worse than when breaking ties arbitrarily.

4. RESULTS ON BROADCAST NEWS 1999

We have verified the modifications of the ROVER algorithm on the 1999 Broadcast News evaluation test set. The focus of this evaluation was on 10xRT large vocabulary continuous speech recognizers. Table 4 summarizes the official results of the individual recognizers and of the reference ROVER[2] [4].

10x RT					50x RT	unlimited	
LIMSI	BBN	IBM	NIST BBN	CMU	**ROVER**	IBM	LIMSI
17.1%	17.3%	17.6%	24.6%	26.3%	**14.4%**	15.0%	15.9%

Table 4: Official word error rates for the 1999 BN evaluation.

Original ROVER achieves a relative word error reduction of 16% with respect to the best single recognizer when used to combine the five 10xRT recognizers in alphabetical order. Note in particular that ROVER achieves significantly better results and that it is faster than any of the unlimited recognizers (each one

[1] Mangu et al. do not report sentence errors [3].

[2] This year, NIST also used normalizing/filtering prior to combination.

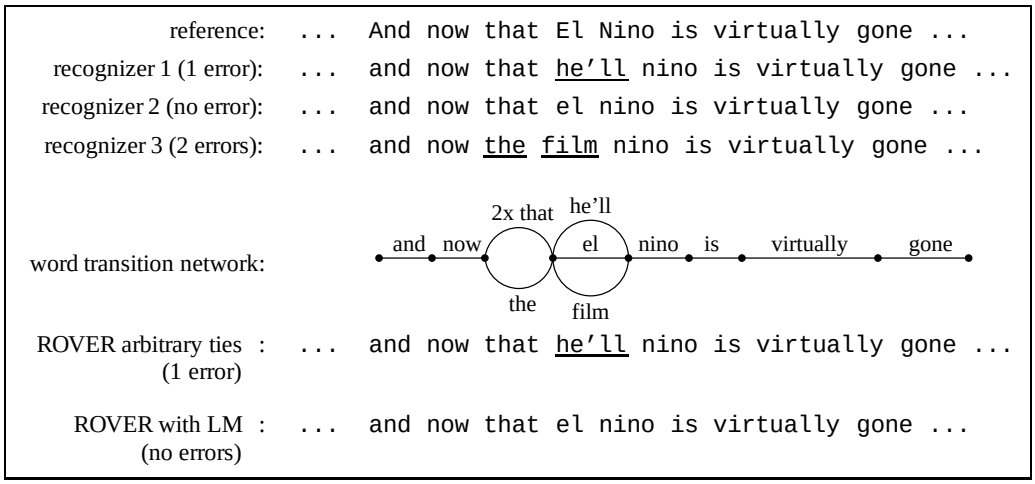

Figure 4: Example of ROVER WTN and result (wrong words are underlined). With the standard ROVER, even though one recognizer has the correct answer it may not be selected since the 3-way tie (`he'll`, `el` and `film`) is broken arbitrarily. With help of the LM, the correct word is selected.

needs in fact more than 50 times real-time). This may indicate a new direction for feature research in speech recognition: developing several fast recognizers and combining them may lead to better performance than one very complicated one.

The difference in the word error rates suggests combining only the three best recognizers. When these three recognizers are combined and a LM is used to break ties we achieve a word error of 13.6% in 30xRT. This is a 5.6% relative improvement with respect to the alphabetical ROVER (14.4% werr, 50xRT) and about 20% relative improvement with respect to the best individual recognizer (17.1% werr, 10xRT). Therefore, this approach seems to be of particular interest for improving the recognition performance by combining a small number of relatively fast systems. Table 5 summarizes the results when two to five 10xRT recognizers are combined.

number of combined systems:	2	3	4	5
arbitrary ties	18.9%	14.3%	14.1%	14.1%
arbitrary ties + LM	15.2%	**13.6%**	13.8%	14.0%
relative improvement	-11.1%	**-20.5%**	-19.3%	-18.1%

Table 5: 1999 Broadcast News test set word error rates when using LM information compared to breaking ties arbitrarily. The relative improvement is indicated with respect to the best single recognizer (17.1% werr).

The order of combination should be determined using the performance of each 1999 recognizer on the previous year's test set (Broadcast News 1998), but this information was not available for all recognizers at the time of writing this paper, so the actual word errors on the 1999 test set were used. However, only minor differences in the results with respect to the ordering of the recognizers are observed. We combined the three best recognizers in all possible orders: the average word error rate was 13.67% and the maximum word error rate was 13.74% (inverse order of the three best recognizers). Figure 4 gives an example of the functioning of ROVER.

5. CONCLUSION

This paper gives a detailed analysis of the behavior of the ROVER voting scheme on the 1998 and 1999 Broadcast News evaluation set. Our experiments indicate that it may hurt performance if too many systems are combined, in particular those with the highest word error rates. Additional improvement can be obtained by filtering/normalizing the outputs of the different speech recognizers prior to combination by ROVER.

Another important result of this work is the fact that language model information can be advantageously used to break ties, in particular when combining only a few recognizers outputs. Using this technique, we have consistently achieved a relative word error reduction of about 5% with respect to the original ROVER algorithm on the quite complicated Broadcast News recognition tasks.

REFERENCES

[1] J. G. Fiscus. A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–354, 1997.

[2] V. Goel and W. J. Byrne. Minimum Bayes-risk automatic speech recognition. *Computer Speech And Language*, 14(2):115–135, 2000.

[3] L. Mangu, E. Brill, and A. Stolcke. Finding consensus among words: Lattice-based word error minimization. In *Eurospeech*, pages 495–498, 1999.

[4] D. S. Pallett, J. G. Fiscus, and J. S. Garofolo. 1999 broadcast news benchmark test results. In *DARPA Broadcast News Workshop, Washington*, May 2000.

[5] D. S. Pallett, J. G. Fiscus, J. S. Garofolo, A. Martin, and M. Przybocki. 1998 broadcast news benchmark test results: English and non-English word error rate performance measures. In *DARPA Broadcast News Workshop, Hernon, VA*, Feb. 1999.

[6] A. Stolcke, Y. König, and M. Weintraub. Explicit word error minimization in n-best list rescoring. In *Eurospeech*, pages 163–165, 1997.