

SEGMENTAL MINIMUM BAYES-RISK ASR VOTING STRATEGIES

Vaibhava Goel

Shankar Kumar

William Byrne *

Center for Language and Speech Processing
The Johns Hopkins University
Baltimore, MD 21218
{vgoel, skumar, byrne}@clsp.jhu.edu

ABSTRACT

ROVER [1] and its successor voting procedures have been shown to be quite effective in reducing the recognition word error rate (WER). The success of these methods has been attributed to their minimum Bayes-risk (MBR) nature: they produce the hypothesis with the least expected word error. In this paper we develop a general procedure within the MBR framework, called segmental MBR recognition, that encompasses current voting techniques and allows further extensions that yield lower expected WER. It also allows incorporation of loss functions other than the WER. We present a derivation of voting procedure of N-best ROVER as an instance of segmental MBR recognition. We then present an extension, called e-ROVER, that alleviates some of the restrictions of N-best ROVER by better approximating the WER. e-ROVER is compared with N-best ROVER on multi-lingual acoustic modeling task and is shown to yield modest yet significant and easily obtained improvements.

1. INTRODUCTION

Voting techniques that combine outputs of multiple recognizers or multiple outputs of a single recognizer have been shown to be useful in reducing the recognition word error rate (WER). First in this series was NIST's ROVER [1] method that combines single hypotheses from multiple ASR systems. Recently introduced voting methods have generalized this to combine multiple recognition hypotheses, contained in N-best lists or lattices, produced by individual [2, 3] and multiple recognizers [4, 5].

Voting procedures start by producing a simultaneous alignment of all the hypotheses. This alignment is stored in a structure called word transition network (WTN) [1] or confusion network [3]. An example WTN is shown in Figure 1. Words that align with each other are grouped together in a correspondence set or bin; {OH, O, !NULL} is one correspondence set in the WTN of Figure 1. A confidence score is then determined for each word and the most confident word is selected from each correspondence set. These selected words are concatenated to produce a hypothesis that forms the output of these systems. Various voting methods essentially differ in the way they produce the WTN and obtain word confidence scores.

*This work was supported by the National Science Foundation under Grant No. #IIS-9810517. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or The Johns Hopkins University.

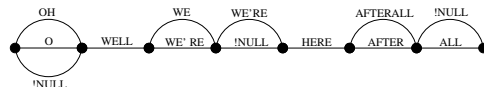


Figure 1: An example word transition network.

The success of voting methods has been attributed to their minimum Bayes-risk (minimum expected error) nature [3, 4, 5, 6, 7, 8]: the cost of the alignment specified by the WTN approximates the word error rate, and the process of selecting most confident words from correspondence sets amounts to picking the hypothesis that has the least expected error under that approximation to WER.

Our intent in this paper is to demonstrate that voting procedures are specific instances of a general procedure which we call *segmental minimum Bayes-risk recognition*. This formulation leads to an improved understanding of the constraints of voting procedures. It also allows us to develop extensions of these procedures that provide a better approximation to the word error rate and that are even applicable to loss functions other than WER.

We start by presenting the framework of segmental MBR recognition. We then show how one voting method, N-best ROVER [4, 5], can be derived as a limiting case of segmental MBR recognition. An extension to N-best ROVER, called e-ROVER, is then presented. e-ROVER is compared with N-best ROVER on the task of multi-lingual language independent acoustic modeling [9].

2. SEGMENTAL MINIMUM BAYES-RISK RECOGNITION

Let A be the acoustic utterance to be recognized. Let \mathcal{W}_e be the set of recognition hypotheses that are likely given A , i.e. word strings that have a non-zero probability $P(W|A)$. Let \mathcal{W}_h be the set of word strings from which the recognizer selects its answer; in general, \mathcal{W}_e could be different from \mathcal{W}_h . We call the former set the *evidence space* and the latter the *hypothesis space*. Let $l(W, W')$ be a loss function that measures the error incurred in hypothesizing W' when A is produced by the evidence word string W . The minimum Bayes-risk (MBR) recognizer that yields the least expected loss under the joint distribution $P(W, A)$ is

$$\delta(A) = \operatorname{argmin}_{W' \in \mathcal{W}_h} \sum_{W \in \mathcal{W}_e} l(W, W') P(W|A). \quad (1)$$

A simplification of the MBR recognizer of Equation 1 can be derived by segmenting the hypothesis and evidence spaces as follows. Let each string in the evidence space \mathcal{W}_e be *uniquely* segmented into M sub-strings of zero or more words. This gives rise to M sets of sub-strings, called *evidence segment sets*. Let \mathcal{W}_e^i denote the i^{th} evidence segment set. Similarly, let each word string in the hypothesis space \mathcal{W}_h be segmented into M sub-strings of zero or more words, giving rise to M *hypothesis segment sets*. Let \mathcal{W}_h^i denote the i^{th} hypothesis segment set. Furthermore, let the hypothesis space be “closed” under concatenation of strings from the hypothesis segment sets

$$\mathcal{W}_h = \mathcal{W}_h^1 \cdot \mathcal{W}_h^2 \cdot \dots \cdot \mathcal{W}_h^M. \quad (2)$$

We define a posterior probability $P(W^i|A)$ for sub-strings W^i of the evidence segment sets \mathcal{W}_e^i as

$$P(W^i|A) = \sum_{X \in \mathcal{W}_e: X^i = W^i} P(X|A). \quad (3)$$

This is the marginal probability of W^i obtained by summing over all those evidence set word strings that contribute sub-string W^i to the i^{th} evidence segment set.

Assume that the segmentation of the spaces is such that the total loss between hypothesis and evidence word strings can be computed (or approximated) by a sum of the losses between their segments

$$l(W, W') = \sum_{i=1}^M l^i(W^i, W'^i), \quad (4)$$

where the superscript i on the word strings denotes their i^{th} sub-string segment, and on the loss function it denotes the loss function defined over the i^{th} segment set. Then, the MBR recognizer of Equation 1 can be implemented as a concatenation of M *segmental MBR recognizers*

$$\delta(A) = \delta^1(A) \cdot \delta^2(A) \cdot \dots \cdot \delta^M(A), \quad (5)$$

where

$$\delta^i(A) = \operatorname{argmin}_{w'^i \in \mathcal{W}_h^i} \sum_{w^i \in \mathcal{W}_e^i} l^i(w^i, w'^i) P(w^i|A). \quad (6)$$

We note that while the utterance level MBR recognizer is implemented as a sequence of independent segmental MBR recognizers on hypothesis and evidence space segments, the acoustic data is not segmented at all. Also, there is no assumption of linguistic independence between word strings belonging to adjacent segments; the language model spans across segments and could even be applied at the utterance level.

Also worth noting is that segmental MBR recognition does not provide a procedure for selecting the hypothesis and evidence segment sets; it only specifies the constraints that these sets need to obey. The construction of segment sets therefore remains a design problem that needs to be addressed in an application specific manner. On a related note, suppose that for a given design of

hypothesis and evidence segment sets, the assumption of Equation 4 is not satisfied. We can then define an *induced* utterance level loss function for that segmentation

$$l_I(W, W') = \sum_{i=1}^M l^i(W^i, W'^i). \quad (7)$$

Clearly, the segmental MBR recognizers of Equation 6 are equivalent to an utterance level MBR recognizer under the loss function l_I . Their performance under the desired loss function l would depend on how well l_I approximates l .

3. VOTING ON SEGMENTS, N-BEST ROVER, AND EXTENDED-ROVER

A special case of segmental MBR recognition arises when each evidence segment set contains at most one word from each evidence word string; each hypothesis segment set contains at most one word from each hypothesis word string; and there is a 0/1 loss function on the segment sets

$$l_{0/1}(w^i, w'^i) = \begin{cases} 0 & \text{if } w^i = w'^i \\ 1 & \text{otherwise.} \end{cases} \quad (8)$$

We are using lower case w^i to indicate that the members of the segment sets are words (or NULL). Under these conditions the segmental MBR recognizer of Equation 6 becomes

$$\delta(A^i) = \operatorname{argmax}_{w'^i \in \mathcal{W}_h^i} P(w'^i|A), \quad (9)$$

where,

$$P(w'^i|A) = \sum_{w^i \in \mathcal{W}_e^i: w^i = w'^i} P(w^i|A). \quad (10)$$

Equation 9 is exactly the maximum a-posteriori decision on each hypothesis segment set. For each word a posterior probability is computed based on the evidence space and then the word with highest posterior probability is selected. We call this *segmental MBR voting*.

It is easily seen that N-best ROVER is an instance of segmental MBR voting. Let N_s be the number of systems being combined. Let $P_k(W|A)$ be the distribution of the k^{th} system, restricted to its N-best list. A single distribution, $P(W|A)$, is generated by taking a convex combination of $P_k(W|A)$. The N-best ROVER procedure can be summarized in the following steps.

1. Construct a WTN from the union of N-best outputs of N_s systems.
2. Using the distribution $P(W|A)$ and the WTN, compute a posterior probability according to Equation 3 for each word in each correspondence set.
3. From each correspondence set, select the word with highest posterior probability. Concatenate these words to produce the final hypothesis.

Clearly, the evidence space in N-best ROVER is the union of N-best lists. The distribution over this evidence space is

$P(W|A)$. The correspondence sets play the role of both the evidence and the hypothesis segment sets, and the hypothesis space is the set of all the paths that are contained in the WTN. The induced utterance level loss function in N-best ROVER is

$$l_R(W, W') = \sum_{i=1}^M l_{0/1}(w^i, w'^i). \quad (11)$$

Since the WTN is constructed to get a good simultaneous alignment between hypotheses, l_R approximates the word error rate. Hence, N-best ROVER is a segmental MBR voting procedure under an approximate WER loss function.

Having at most one word in each segment set restricts segmental MBR voting procedure in that it can not incorporate dependencies in the loss function that span multiple words. This limitation may be of significance in tasks that are sensitive to multiple word dependencies, such as detection of phrases. In fact, even for the task of minimizing word error rate, which is the main objective of N-best ROVER, this restriction is of significance. In some cases it may not be possible to find a simultaneous word level alignment for which the sentence level loss equals the Levenshtein distance between any pair of sentences, as shown by the following three sentences.

OH	WELL	WE	!NULL
O	WELL	WE'RE	!NULL
!NULL	WELL	WE	WE'RE

We now specify a procedure that allows two or more consecutive words in each correspondence set. We first define a process of *joining* two consecutive correspondence sets. In joining two correspondence sets we replace those two sets by one *expanded* set that contains all the paths from the original pair of sets. This is illustrated in Figure 2.

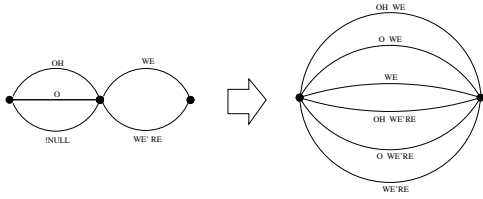


Figure 2: Joining two correspondence sets.

We now introduce the e-ROVER loss function. In a WTN if we join any two consecutive correspondence sets, say absorb sets $m + 1$ into set m , and use the Levenshtein distance loss function L on the expanded set m , then the utterance level loss function,

$$l_{eR}(W, W') = \sum_{i=1, i \neq m, i \neq m+1}^M l_{0/1}(w^i, w'^i) + L(W^m, W'^m), \quad (12)$$

satisfies the inequalities

$$L(W, W') \leq l_{eR}(W, W') \leq l_R(W, W').$$

These inequalities follow as a direct consequence of the definition of Levenshtein distance.

The joining procedure can be carried out many times to yield successively better approximations to the Levenshtein distance. The WTN obtained after each joining operation specifies a new segmentation of the evidence and hypothesis spaces. It is important to note that only the segmentation of the hypothesis and evidence spaces changes with joining operation; the actual spaces remain unchanged.

The size of each expanded set grows exponentially with the number of joining operations, making Equation 6 progressively more difficult to implement. Thus, even though each joining operation improves word error rate approximation, it is important to select the sets to be joined carefully. We use a heuristic method to determine sets that are joined:

1. Construct a WTN, as in N-best ROVER, using N hypotheses each from N_s systems.
2. Determine the posterior probability of words in correspondence sets, according to Equations 3 and 10, as in N-best ROVER.
3. “Pinch” correspondence sets in which the largest value of the posterior probability is above a *pinching threshold*. Join all adjacent unpinched correspondence sets.
4. Use the Levenshtein distance loss function within each expanded set.

The procedure of pinching and expanding the correspondence sets is shown in Figure 3. Hypotheses in e-ROVER are formed sequentially according to Equations 5 and 6.

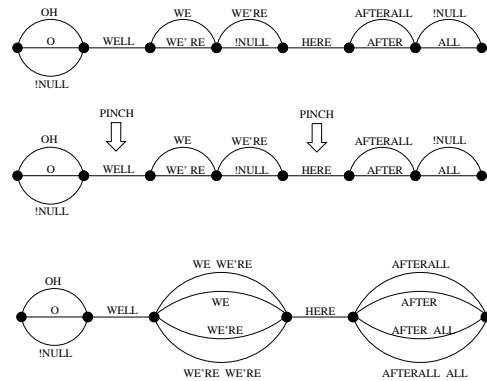


Figure 3: e-ROVER WTN construction.

As noted above, the hypothesis and evidence spaces in e-ROVER are identical to those in N-best ROVER. However, the loss function in e-ROVER provides a better approximation to the word error rate due to the improved segmentation. Since they are both instantiations of Equation 5, e-ROVER would be reasonably expected to yield a lower word error rate than N-best ROVER.

4. PRELIMINARY RESULTS

Our preliminary comparisons of N-best ROVER with e-ROVER were performed on the multi-lingual language independent acoustic modeling task [9]. Czech recognition outputs from three systems were combined : a triphone system trained on one hour of Czech voice of America (Cz-VOA) database (Sys1); a

triphone system trained on 72 hrs. of English and then adapted to one hour of Czech (Sys2); and Sys1 output rescored with Sys2 models. The test set consisted of 748 held out utterances from Cz-VOA broadcast. 250 hypotheses were taken from each system along with their distributions restricted to these 250-best lists. The baselines (MAP hypotheses) in these systems had error rates of 29.42%, 35.24%, and 29.22%, respectively.

N-best ROVER yields an absolute improvement of 3.28% over the 29.22% baseline. Its comparison with e-ROVER is shown in Figure 4. The top panel in this figure shows the fraction of the pinched sets as a function of the pinching threshold. A threshold of 0.0 pinches all the sets - equivalent to N-best ROVER - while any threshold above 1.0 results in no pinching at all. At higher pinching thresholds (i.e. for fewer pinched sets) the size of expanded sets grew beyond computational capacities and a likelihood based pruning was applied within the expanded sets. This resulted in a smaller effective hypothesis space at higher thresholds. The bottom panel in Figure 4 shows the effect of pinching threshold on the word error performance of e-ROVER. We note that all thresholds result in better than N-best ROVER performance. The threshold of 1.0 yields the best performance of 0.56% absolute improvement over N-best ROVER and hence a total of 3.84% absolute over the baseline error rate of 29.22%. We see a degradation in performance for thresholds larger than 1.0, owing to the pruning of the expanded sets.

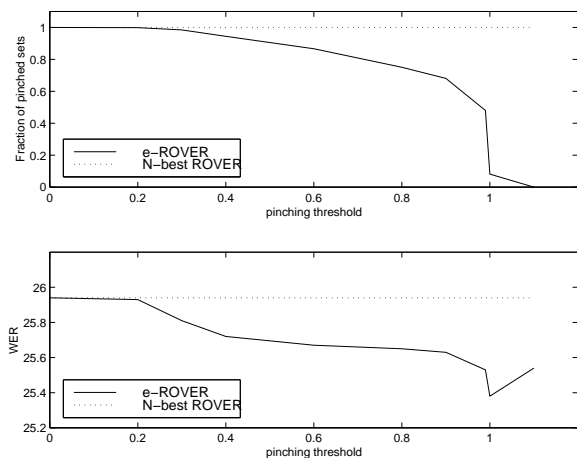


Figure 4: Fraction of pinched correspondence sets and word error rate performance of e-ROVER as a function of pinching threshold.

5. DISCUSSION AND CONCLUSIONS

In this paper we have presented segmental MBR recognition and shown how the voting method of N-best ROVER can be derived as a special case of segmental MBR recognition. Consequent to this formulation we derived e-ROVER as an extension to N-best ROVER. e-ROVER provides an improved approximation to word error rate and on one practical task yields small but significant and easily obtained error rate reduction.

While e-ROVER allows incorporation of multiple consecutive words in each correspondence set, the loss function on each set is still the Levenshtein distance. An immediate extension would be to use other interesting measures of distance between words

and word strings; for example $l(\text{HELLO}, \text{LOW})$ could be larger than $l(\text{HELLO}, \text{YO!})$. This would allow for many useful sentence level loss functions without adding too much complexity to the hypothesis selection process.

One of the primary shortcomings of e-ROVER, as it is described here, is the ad-hoc pinching of correspondence sets. Pinching should be driven by the overall Bayes-risk and not by posterior probabilities alone. We are currently investigating more rigorously formulated WTN pinching procedures.

6. REFERENCES

1. J. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347–354, Santa Barbara, CA, 1997.
2. L. Mangu, E. Brill, and A. Stolcke. Searching for consensus to improve recognition output. In *Proceedings of the 9th Hub-5 Conversational Speech Recognition Workshop*, Linticum Heights, MD, 1998.
3. L. Mangu, E. Brill, and A. Stolcke. Finding consensus among words: lattice-based word error minimization. *Eurospeech-99*, pp. 495–498, Budapest, Hungary, 1999.
4. G. Evermann and P. C. Woodland. Posterior Probability Decoding, Confidence Estimation and System Combination. In *Proceedings of the Speech Transcription Workshop*, May 16–19, University of Maryland, College Park, MD, 2000.
5. V. Goel and W. Byrne. Applications of Minimum Bayes-Risk Decoding to LVCSR. In *Proceedings of the Speech Transcription Workshop*, May 16–19, University of Maryland, College Park, MD, 2000.
6. V. Goel and W. Byrne (advisor). Thesis proposal for the dept. of Biomedical Engineering. April 30, 1999. Available upon request.
7. V. Goel and W. Byrne. Task dependent loss functions in speech recognition: A^* search over recognition lattices. *Eurospeech-99*, pp. 1243–1246, Budapest, Hungary, 1999.
8. V. Goel and W. Byrne. Minimum Bayes-risk automatic speech recognition. *Computer Speech and Language*, Vol. 14(2), pp. 115–135, 2000.
9. W. Byrne, et.al. Towards language independent acoustic modeling. *ICASSP00*, pp. 1029–1032, Istanbul, Turkey, 2000.

SEGMENTAL MINIMUM BAYES-RISK ASR VOTING STRATEGIES

*Vaibhava Goel, Shankar Kumar and William Byrne*¹

Center for Language and Speech Processing

The Johns Hopkins University

Baltimore, MD 21218

{vgoel, skumar, byrne}@clsp.jhu.edu

ROVER [1] and its successor voting procedures have been shown to be quite effective in reducing the recognition word error rate (WER). The success of these methods has been attributed to their minimum Bayes-risk (MBR) nature: they produce the hypothesis with the least expected word error. In this paper we develop a general procedure within the MBR framework, called segmental MBR recognition, that encompasses current voting techniques and allows further extensions that yield lower expected WER. It also allows incorporation of loss functions other than the WER. We present a derivation of voting procedure of N-best ROVER as an instance of segmental MBR recognition. We then present an extension, called e-ROVER, that alleviates some of the restrictions of N-best ROVER by better approximating the WER. e-ROVER is compared with N-best ROVER on multi-lingual acoustic modeling task and is shown to yield modest yet significant and easily obtained improvements.