I. Introduction

- Minimum Bayes Risk (MBR) decoders improve upon MAP decoders by directly optimizing loss function of interest: Word Error Rate
- MBR decoding is expensive when the search spaces are large
- Segmental MBR (SMBR) decoding breaks the single utterance-level MBR decoder into a sequence of simpler search problems.

To do this, the N-best lists or lattices need to be segmented.
- We present: A new lattice segmentation strategy based on a risk criterion

II. Segmental Minimum Bayes-Risk Decoding

- Minimum Bayes-risk (MBR) decoder on an utterance A:
  \[ \delta(A) = \min_{W \in \mathcal{W}} \sum_{W^1 \in \mathcal{W}} t(W, W^1)p(W^1|A) \]
- MBR Implementations: N-best rescoring, lattice based, A* search and lattice based extended ROVER (e-ROVER)
- A word lattice

Suppose we can segment this word lattice:

This segmentation induces a loss function between any two word strings:

\[ t(W, W^1) = t(W_1, W_1^1) + t(W_2, W_2^1) \]

A good segmentation is such that the true loss function is the same as the induced loss function:

\[ t(W, W^1) = t(W, W^1) \]

Under such a segmentation, MBR decoder reduces to a concatenation of MBR decoders:

\[ \delta(A) = \min_{W \in \mathcal{W}} \sum_{W \in \mathcal{W}} t(W, W)p(W|A) \]

III. Ideal Lattice Segmentation

- Goal: Reduce search space of MBR recognizer
  - Pruning the lattice could result in search errors
  - Segmentation breaks up a single search problem into many simpler search problems
  - Any segmentation restricts string alignments, \( t(W, W^1) \leq \sum_{i=1}^{N} t(W_i, W_i^1) \)
- Therefore, segmentation involves tradeoff between search errors and errors in approximating the loss function

IV. Risk Criterion for Lattice Segmentation

- Total Bayes-Risk of all lattice word strings
  \[ R_T = \sum_{W \in \mathcal{W}} t(W^1, W)p(W^1|A) \]

- \( \delta^* = \delta^* \) is the MAPP string in the lattice
- ML approximation to Total Bayes-Risk
  \[ R_T \approx R_T = \sum_{W \in \mathcal{W}} t(W^1, W)p(W^1|A) \]

- After Lattice Segmentation
  \[ R_T \leq P(W|A) \sum_{W \in \mathcal{W}} t(W, W^1)p(W^1|A) \]

- Modified Segmentation Criterion: Minimize the upper bound on \( R_T \)

V. Levenshtein Alignment via Weighted Finite State Transducers

- Goal: Optimal alignment under Levenshtein distance between \( W \) and \( W^1 \)
- We have developed an efficient Weighted Finite State procedure that computes the Levenshtein alignment between \( W \) and all lattice word strings.

- Algorithmic details in the paper
- The procedure tags every lattice link with an index \( j \) along the best path: \( j \in \{1, 2, \ldots, K\} \)

VI. Risk-Based Lattice Cutting

- Risk-Based Lattice Cutting (RLC)
  - Segment the lattice into \( K \) segments based on the Levenshtein alignment
- Periodic Risk-Based Lattice-Cutting (PLC)
  - Segment lattice into \( K \) segments by choosing node sets at equal periods
  - Higher Period: Better approximation to the Levenshtein distance
  - Higher Period: More Search errors
  - RLC : PLC with a period of 1
  - PLC with Period of 2
  - PLC with Period of 3

VII. Results on SWITCHBOARD

- Johns Hopkins University LVCSR Hub 5 2001 Evaluation System
- Test sets: Swbd-2 portion of 1998 evalset (SWB2), Swbd-1 portion of 2000 evalset (SWB1)

<table>
<thead>
<tr>
<th>Decoding Strategy</th>
<th>WER(%)</th>
<th>MMIE</th>
<th>MAP</th>
<th>MMIE - MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-best rescoring</td>
<td>40.4</td>
<td>25.0</td>
<td>25.6</td>
<td></td>
</tr>
<tr>
<td>Entire lattice</td>
<td>41.0</td>
<td>25.9</td>
<td>26.1</td>
<td></td>
</tr>
<tr>
<td>PLC</td>
<td>40.1</td>
<td>25.4</td>
<td>24.7</td>
<td></td>
</tr>
<tr>
<td>Entire lattice</td>
<td>41.0</td>
<td>25.5</td>
<td>25.5</td>
<td></td>
</tr>
<tr>
<td>PLC</td>
<td>40.0</td>
<td>25.4</td>
<td>25.2</td>
<td></td>
</tr>
<tr>
<td>Entire lattice</td>
<td>40.5</td>
<td>25.7</td>
<td>24.9</td>
<td></td>
</tr>
<tr>
<td>PLC</td>
<td>39.5</td>
<td>25.3</td>
<td>24.3</td>
<td></td>
</tr>
</tbody>
</table>

- SMBR decoding on segmented lattices performs better than MAP or MBR decoding on unsegmented lattices
- SMBR decoding performs better under PLC than under RLC

VIII. An Application to Multiple-System Lattice Combination

- Investigate ASR System Combination Strategies using Lattices
  - Johns Hopkins University RT-02 LVCSR Evaluation System
  - MMR: 40.9, 39.6, 39.6
  - MMIE: 40.4, 39.5, 39.5
  - PLC: 39.9, 39.9, 39.9
  - Segmental Lattice: 39.8, 39.9, 39.9

- SMBR decoding is better than simply intersecting lattices and rescoring
- Adding posteriors of hypotheses over sub-lattices is better than multiplying them (Using a conditional independence assumption)

IX. Conclusions

- A lattice cutting procedure based on a risk criterion
  - Segment the lattice wrt the MAP hypothesis
  - No time information or likelihoods required from the lattice
- Performance of MBR procedures improves when applied to lattice cuts
- PLC procedure performs better than RLC procedure
- Proper tradeoff between Levenshtein distance approximation and search errors is crucial
- Applications of Lattice Segmentation
  - Multiple-system lattice combination via sub-lattice combination
  - Confidence estimation for recognized word strings