GENERALIZED DRIVEN DECODING FOR SPEECH RECOGNITION SYSTEM COMBINATION

Benjamin Lecouteux, Georges Linares, Yannick Estève, Guillaume Gravier

To cite this version:

Benjamin Lecouteux, Georges Linares, Yannick Estève, Guillaume Gravier. GENERALIZED DRIVEN DECODING FOR SPEECH RECOGNITION SYSTEM COMBINATION. ICASSP, 2008, Las Vegas, United States. hal-02094742

HAL Id: hal-02094742
https://hal.archives-ouvertes.fr/hal-02094742
Submitted on 9 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
GENERALIZED DRIVEN DECODING FOR SPEECH RECOGNITION SYSTEM COMBINATION

Principle of driven decoding algorithm (DDA)

- The DDA based combination:
  - A first recognition pass using an auxiliary ASR system
  - The auxiliary system provides the best hypothesis
  - The transcript drive the main search algorithm
  - DDA is an integrated approach for system combination

Anatomy of the Spectral decoder

- Large vocabulary continuous speech recognition system
- HMM-based acoustic modeling
- Trigram language models
- Search: derived from a A* search algorithm operating on a lattice of phonemes
- Exploration is supervised by the function $F(h_n)$ evaluating the probability of $h_n$ crossing the node $n$:
  \[ F(h_n) = g(h_n) + p(h_n) \]

DDA step 1: on-demand synchronization

- Spectral speech recognition system generates hypotheses as the phoneme-lattice is explored
- $A^*$ is an asynchronous decoder
- Hypotheses are extended or left according to $F()$
- Leave a path leads to backtracking
- DDA synchronizes the current hypothesis and the auxiliary transcript
- Synchronization by fast DTW algorithm

DDA step 2: transcript to hypothesis matching score

- Linguistic probabilities are modified using the following rescoring rule:
  \[ L(w_j|w_{j-1}, w_{j-2}, \ldots, w_1) = P(w_j|w_{j-1}, w_{j-2}, \ldots, w_1) \cdot \hat{o}(w_j) \]  
  where $\hat{o}(w_j)$ is the resulting linguistic score, $P(w_j|w_{j-1}, w_{j-2}, \ldots, w_1)$ the initial probability, $\gamma$ an empirical fudge factor and $\hat{o}(w_j)$ is the confidence score of $w_j$ given by:
  \[ \hat{o}(w_j) = \begin{cases} \hat{o}(w_j) & \text{if } \gamma > 0 \\ \gamma \cdot \hat{o}(w_j) & \text{if } \gamma < 0 \end{cases} \]  
- $\gamma$ is the analysis window size reported by the edit distance ($\gamma = 4$) and $\hat{o}(w_j)$ posteriors from word $w_j$ of the auxiliary system.

Experimental framework

- The LIA system
  - System involved in the ESTER evaluation campaign
  - Speech decoder
  - Alphabet-based segmenter
  - 6k lexicon, 24M triphones estimated on about 2000 hours of speech
  - 2-decoding pass (LILK adaptation)
  - SoftRNN on a standard desktop computer

- The LIUM system
  - Based on the CMU Sphinx 3.2 decoder (beam search algorithm)
  - English vocal tract resonance process
  - Context-dependent acoustic models trained on TPR materials
  - SAT-based adaptation
  - The entire system runs under 12GoT

- The IRISA system
  - Based on word synchronous beam search algorithm
  - HMM-acoustic modeling and g-synthetic linguistic models with a vocabulary of 246k words
  - The system operates in four steps
    - 1) Context-independent acoustic models with a triphone LM
    - 2) Search graph is generated with a trigram LM and context-dependent models
    - 3) MLU-synthetic model adaptation
    - 4) Consensus decoding is applied to the 1000-best sentence hypotheses

Baseline results

- The Driven Decoding Algorithm (DDA) is used here during the second pass

The Driven Decoding Algorithm

Confusion network driven decoding

- Two-Level ROVER-DDA combination
  - Relies on a first merging step where all auxiliary transcripts are merged
  - We use ROVER for merging LIUM and IRISA system outputs
  - The word confidence scores of the output are computed by averaging the confidence scores of words in each single system output
  - The resulting transcript is then used as an auxiliary hypothesis

Integrated DDA-based combination

- All auxiliary systems outputs are submitted
- For each of them, a matching score is computed according to independent transcript-to-hypothesis synchronization
- All linguistic scores are merged by the log-linear combination extended to n-systems:
  \[ L_i(w_j) = L_i(w_j) \cdot \frac{1}{N} \sum_{j=1}^{N} L_j(w_j) \]  
  where $L_i$ is the averaged $L_i$ as defined in equation 2, $w_j$ are the posterior probabilities provided by the system $i$ and $N$ the number of auxiliary systems.

Multi system combination

Two-Level ROVER-DDA combination

<table>
<thead>
<tr>
<th>System</th>
<th>F Inter</th>
<th>F Info</th>
<th>RFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUM</td>
<td>18.5</td>
<td>18.9</td>
<td>25.6</td>
</tr>
<tr>
<td>DDA-LUM-P1</td>
<td>17.9</td>
<td>18.1</td>
<td>22.4</td>
</tr>
<tr>
<td>DDA-LUM-P2</td>
<td>18.5</td>
<td>17.8</td>
<td>21.5</td>
</tr>
<tr>
<td>DDA-RCN-LUM-P1</td>
<td>17.7</td>
<td>18.1</td>
<td>23.3</td>
</tr>
<tr>
<td>DDA-RCN-LUM-P2</td>
<td>17.2</td>
<td>17.8</td>
<td>21.5</td>
</tr>
</tbody>
</table>

Table: Word error rates for confusion network driven decoding (DA-WCR), according to the decoding pass. Results are compared to the one of the best single system (LUM) and to the best one-best DDA system (DDA-LUM)