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.
The Driven Decoding Algorithm

**Principle of driven decoding algorithm (DDA)**
- The DDA based combination:
  - A first recognition pass using an auxiliary ASR system
  - Auxiliary system provides the one-best hypothesis $h_{aux}$
  - The auxiliary transcript drives the main search algorithm
  - DDA is an integrated approach for system combination

- Transcripts drive $A^*$ decoding:
  - $A^*$ search is synchronized to the transcript
  - Linguistic probabilities are dynamically rescored
  - Rescoring is based on posterior probabilities

- **Anatomy of the Speeral 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 $F(h_a)$ evaluating the probability of $h_a$ crossing the node $n$: 
    $$ F(h_a) = g(h_a) + p(h_a) $$

- **DDA step 1: on-demand synchronization**
  - Speeral 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:
    $$ L(w_i, w_{i-1}) = P(w_i | w_{i-1}) - \gamma \cdot \rho(w_i) $$
    where $\gamma$ is the analysis window size reported by the edit distance ($\sim 4$) and $\rho(w_i)$ posterior from word $w_i$ of the auxiliary system.

**Experimental framework**
- **The LIA system**
  - System involved in the ESTER evaluation campaign
  - Speech decoder
  - Alphabet-based segmentation
  - 66k emission, 2M bigrams estimated on about 200k words
  - 2 decoding-pass (LILM adaptation)
  - Switch on a standard desktop computer
- **The LIUM system**
  - Based on the CMU Sphinx-3.5 decoder (beam search algorithm)
  - English word lattice rescoring process
  - Context-dependent acoustic models trained on Estier materials
  - IAT-based adaptation
  - The entire process runs under 120RT
- **The IRISA system**
  - Based on a word synchronous beam search algorithm
  - HMM acoustic modeling and h-gram linguistic models with a vocabulary of 28k words
  - The system operates in four steps:
    1. Context-independent acoustic models with a trigram LM
    2. The graph is rescoring with a trigram LM and context-dependent models
    3. Multi-LM speaker adaptation
    4. Consensus decoding is applied to the 1000-best sentence hypotheses

**Baseline results**
- ASR systems are assessed on 3 hours of radio broadcast from ESTER corpus
- The Driven Decoding Algorithm (DDA) is used here during the second pass

**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:
  $$ L(w_i, w_{i-1}) = P(w_i | w_{i-1}) - \gamma \cdot \rho(w_i) $$
  where $\gamma$ is the averaged $\gamma$ as defined in equation 2, $\rho(w_i)$ are the posteriori provided by the system $i$ and $N$ the number of auxiliary systems.

**Results**

<table>
<thead>
<tr>
<th>System</th>
<th>F Inter</th>
<th>F Info</th>
<th>RFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIUM</td>
<td>18.5</td>
<td>18.9</td>
<td>25.8</td>
</tr>
<tr>
<td>DDA-LIUM-P1</td>
<td>17.8</td>
<td>18.1</td>
<td>22.4</td>
</tr>
<tr>
<td>DDA-LIUM-P2</td>
<td>17.2</td>
<td>17.8</td>
<td>21.5</td>
</tr>
<tr>
<td>DDA-LIUM-LUM-P1</td>
<td>17.7</td>
<td>18.1</td>
<td>23.3</td>
</tr>
<tr>
<td>DDA-LIUM-LUM-P2</td>
<td>17.2</td>
<td>17.8</td>
<td>21.5</td>
</tr>
</tbody>
</table>

**Analysis of DDA by comparison to ROVER and Oracle measures**
- DDA-driven decoding improves the primary system but performance are very close to the one obtained with the more simple one-best driven decoding
- By using DDA-based cross-system combination and a final ROVER pass, we obtained a global absolute gain of about 3.5% WER (15.7% relative gain)

This research is supported by the ANR (Agence Nationale de la Recherche), EPAC project ANR-06-MDCA-006