Improved Context Integration for Robust Speech Recognition in Conversational Systems

20.5.2006

Georg Stemmer

University Erlangen-Nürnberg

now with Siemens AG, Corporate Technology
Overview

1. influence of variation of spoken language on a speech recognizer

2. integration of context to model variability
   - feature extraction
   - acoustic models
   - language model

3. combination of approaches

4. future work

5. conclusion
The \textit{Evar} Spoken Dialogue System

- \textit{Evar} is a spoken dialogue system for train-timetable information developed at the University Erlangen-Nürnberg
- it has been connected to the public phone network for several years
- a sample dialogue is (translated from German):

\begin{verbatim}
Evar: what information do you need?
user: I’d like to go to Frankfurt next week.
Evar: where do you want to leave?
user: from Erlangen.
Evar: on which date do you want to go?
user: on Monday next week if possible.
Evar: what time do you want to leave?
user: I thought around two p.m.
Evar: Your connection is ... (lengthy timetable information)
user: (hangs up)
\end{verbatim}
Influence of Variation on a Speech Recognizer

Variability in the Spoken Input of Evar

- with the Evar system a corpus of \(\approx 12\) hours of telephone speech has been collected

- the telephone calls have been assigned labels to characterize speakers and acoustic conditions

<table>
<thead>
<tr>
<th>label</th>
<th>number of calls</th>
<th>[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>1589</td>
<td>85.2</td>
</tr>
<tr>
<td>female</td>
<td>276</td>
<td>14.8</td>
</tr>
<tr>
<td>children</td>
<td>88</td>
<td>4.7</td>
</tr>
<tr>
<td>elderly</td>
<td>10</td>
<td>0.5</td>
</tr>
<tr>
<td>non-native</td>
<td>93</td>
<td>5.0</td>
</tr>
<tr>
<td>dialect</td>
<td>89</td>
<td>4.8</td>
</tr>
<tr>
<td>low-volume</td>
<td>442</td>
<td>23.7</td>
</tr>
<tr>
<td>bad acoustics</td>
<td>158</td>
<td>8.5</td>
</tr>
<tr>
<td>labeled</td>
<td>1865</td>
<td>100</td>
</tr>
<tr>
<td>all</td>
<td>1968</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Influence of Variation on a Speech Recognizer

Influence of Variation on Recognition Results

- train a speech recognizer on the \textit{Evar} corpus
  - standard semi-continuous HMM-based speech recognizer
  - codebook with 500 full-covariance Gaussian densities
  - 24-dimensional feature vector (MFCCs + energy + derivatives)
  - feature normalization:
    band-pass energy normalization and cepstral mean subtraction
  - bigram language model in the beam search phase
  - 4-gram language model in a second \textit{A*}-search phase

- partitioning of the \textit{Evar} data:

\begin{center}
\begin{tabular}{|l|c|c|}
\hline
\textbf{task} & \textbf{utterances} & \textbf{words} \\
\hline
training & 8000 & 28117 \\
validation & 2000 & 7009 \\
test & 2500 & 8254 \\
all & 12500 & 43380 \\
\hline
\end{tabular}
\end{center}
Influence of Variation on a Speech Recognizer

**Influence of Variation on Recognition Results**

- measure performance for all labels that occur sufficiently frequent in test data
- word error rate (WER) ranges between 20.4% (male speakers) and 49.4% (children)
Influence of Variation on a Speech Recognizer

**Influence of Variation on Recognition Results – Discussion**

- only 62.0% of the calls are from adult native speakers who are up to 60 years old and spoke loud enough without too much dialect or noise
- WER on these speakers is 18.1%
  → lack of robustness is responsible for a 21% relative increase in WER (at least)
- problem:
  - many different speaker groups / conditions with a large variability
  - each of them does not occur often enough to train a separate set of models
- standard approach: adaptation (MLLR, . . . )
  - difficult to apply in spoken dialogue systems:
    - time-synchronous decoding + very short utterances
- ⇒ here: integrate context to achieve higher robustness against unknown sources of variability
Integration of Context to Model Variability

context can be utilized at several stages of the speech recognition process:

- feature extraction
  - static features represent 16 milliseconds of speech
  - dynamic features are computed from five consecutive feature vectors
    → improve the dynamic features to cover a larger context

- acoustic models
  - the output densities of the acoustic models are context-independent
    → introduce context into the output densities of the HMMs

- language model
  - baseline language model is static
    → adapt the language model to the current dialogue-state
Feature Extraction

- dynamic features are the slope of a regression line that approximates the derivative of the static features

- conventionally they are computed from a fixed time segment (e.g. 5 frames)

⇒ combine several different time-resolutions into one feature vector
Multiple Time-Resolutions: Feature Transformation

- three time-resolutions are combined: 3, 5, 7 frames
- only the dynamic features are transformed, the static features are left unchanged

<table>
<thead>
<tr>
<th>feature transformation</th>
<th>none</th>
<th>PCA</th>
<th>LDA</th>
<th>HDA</th>
<th>baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER [%]</td>
<td>22.8</td>
<td><strong>20.9</strong></td>
<td>22.9</td>
<td>22.4</td>
<td>21.9</td>
</tr>
</tbody>
</table>
Multiple Time-Resolutions: PPCA densities

- Probabilistic PCA (PPCA) densities represent variation in the direction of the eigenvectors which correspond to the smallest eigenvalues as ‘noise’

- The output densities of the HMMs can be more robust when PPCA densities replace the full-covariance Gaussian densities in the codebook of the baseline system

- PPCA densities model the high-dimensional, variance-normalized feature vector

<table>
<thead>
<tr>
<th>feature vector</th>
<th>density type</th>
<th>feature transformation</th>
<th>dimension</th>
<th>WER [%]</th>
<th>rel. impr. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed time-res.</td>
<td>Gauss</td>
<td>none</td>
<td>24</td>
<td>21.9</td>
<td>n/a</td>
</tr>
<tr>
<td>multiple time-res.</td>
<td>Gauss</td>
<td>PCA</td>
<td>24</td>
<td>20.9</td>
<td>4.6</td>
</tr>
<tr>
<td>multiple time-res.</td>
<td>Gauss</td>
<td>none</td>
<td>48</td>
<td>22.8</td>
<td>-4.1</td>
</tr>
<tr>
<td>multiple time-res.</td>
<td>PPCA</td>
<td>none</td>
<td>48→24</td>
<td>20.5</td>
<td>6.4</td>
</tr>
</tbody>
</table>
Integration of Context to Model Variability – Acoustic Models

Context-Dependent Output Densities for HMMs

- output density $b_i$ of a semi-continuous HMM:
  \[
  b_i(x_t) = \sum_m P(m|i) \cdot \mathcal{N}(x_t|\mu_m, \Sigma_m)
  \]

- introduce random variable $l +$ independence assumptions $+$ weighting factor $w$:
  \[
  b_i(x_t|x_{1}^{t-1}) \approx \left[ \sum_m P(m|i) \cdot \mathcal{N}(x_t|\mu_m, \Sigma_m) \right]^w \cdot \left[ \sum_l P(l|i) \cdot P(l|x_{1}^{t-1}) \cdot p(x_t|l) \right]^{1-w}
  \]
Integration of Context to Model Variability – Acoustic Models

Integration of a Phone Recognizer

- define the class-label $l$ to be a cluster of states in a phone recognizer → efficient evaluation of $p(x_t|l) \cdot P(l|x_1^{t-1})$ with the beam search algorithm
- example: *gegen vierzehn Uhr* (around 2 o’clock)
Context-Dependent Output Densities: Results

phone recognizer:

- phone error rate of 39.7% on the test data
- number of clusters is fixed to 20 based on results in previous experiments

word recognizer:

- weighting factor set to $w = 0.5$ in training, optimized on validation set for decoding
- the optimal weight for the validation data is $w = 0.8$
- WER on test set is 20.9% (4.6% rel. impr. over baseline WER 21.9%)
Adaptation of the Language Model to the Current Dialogue-State

- each utterance in the *Evar* corpus corresponds to a certain system prompt

<table>
<thead>
<tr>
<th>prompt-id</th>
<th>utterances</th>
<th>[%]</th>
<th>system prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1527</td>
<td>12.2</td>
<td><em>what information do you need?</em></td>
</tr>
<tr>
<td>2</td>
<td>1186</td>
<td>9.5</td>
<td><em>what time do you want to leave?</em></td>
</tr>
<tr>
<td>6</td>
<td>631</td>
<td>5.1</td>
<td><em>where do you want to go?</em></td>
</tr>
<tr>
<td>9</td>
<td>971</td>
<td>7.8</td>
<td><em>where do you want to leave?</em></td>
</tr>
<tr>
<td>10</td>
<td>1443</td>
<td>11.5</td>
<td><em>on which date do you want to go?</em></td>
</tr>
<tr>
<td>0</td>
<td>6742</td>
<td>53.9</td>
<td>everything else, e.g.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>shall I repeat the connection?</em>,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>please connect to a human operator</em></td>
</tr>
<tr>
<td>0 - 10</td>
<td>12500</td>
<td>100</td>
<td>all</td>
</tr>
</tbody>
</table>

- syntax and semantics of an utterance depend on the system prompts which mark also the different *dialogue-states*
Dialogue-State Dependent Language Model: Approach

- **approach:** train a separate language model for each dialogue-state $d$

- **problem:** for each language model $L_d$ there is only a fraction of the total training data available $\Rightarrow$ precise, but less robust estimation

- **solution:**
  - interpolate dialogue-state dependent language model $L_d$ (precise) with the global dialogue-state independent language model $L$ (robust)
  - we compare linear and rational interpolation
  - the interpolation weights are optimized on the validation data
Integration of Context to Model Variability – Language Model

**Rational Interpolation Between Language Models**

- predictors $\hat{P}^d_i(w|h)$ and $\hat{P}_i(w|h)$ of both language models $L_d$ and $L$ estimate the probability of the current word $w = w_s$ given the last $i$ words in history $h$

- non-linear functions $g_i(h)$ weight the predictors with their estimated reliability

$\Rightarrow$ probability $P^d_i(w | h)$ of a word $w$ given its history $h$ for dialogue-state $d$:

$$ P^d_i(w | h) = \frac{\sum_{i=0}^{n} \rho_i^d \cdot g_i^d(h) \cdot \hat{P}^d_i(w|h) + \varrho \cdot \sum_{i=0}^{n} \rho_i \cdot g_i(h) \cdot \hat{P}_i(w|h)} {\sum_{i=0}^{n} \rho_i^d \cdot g_i^d(h) + \varrho \cdot \sum_{i=0}^{n} \rho_i \cdot g_i(h)} $$

- using weights $\rho_i,\rho_i^d,\varrho$
Integration of Context to Model Variability – Language Model

**Dialogue-State Dependent Language Model: Results**

![Graph showing word error rates and relative improvements in WER for different interpolation methods.]

- **Baseline** WER: 21.9%
- **Without interpolation** WER: 21.7%
- **Linear interpolation** WER: 21.6%
- **Rational interpolation** WER: 21.0%

<table>
<thead>
<tr>
<th>Dialogue-State</th>
<th>WER [%]</th>
<th>Rel. Imp. in WER (rational)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>21.6</td>
<td>0.3%</td>
</tr>
<tr>
<td>1</td>
<td>21.7</td>
<td>-0.3%</td>
</tr>
<tr>
<td>2</td>
<td>21.7</td>
<td>7.6%</td>
</tr>
<tr>
<td>6</td>
<td>21.0</td>
<td>11.2%</td>
</tr>
<tr>
<td>9</td>
<td>21.0</td>
<td>18.3%</td>
</tr>
<tr>
<td>10</td>
<td>21.0</td>
<td>10.8%</td>
</tr>
</tbody>
</table>
Combination of Approaches

- MFCCs + derivatives
  - multiple time-resolution

- PPCA density
  - evaluation

- phone recognizer
  - state scores
  - codebook scores

- Viterbi beam search
  - combine scores

- dialogue manager
  - dialogue-state

- A*-search
  - recognized
  - word chain

 acoustic input

- word graph
Combination of Approaches

Results

- relative improvements are higher for labels with high error rates
- maximum possible reduction for all is $6.4\% + 4.6\% + 4.0\% = 15.0\%$
  $\rightarrow$ redundancy between individual approaches is relatively low
Conclusion and future work

• current speech recognizers are not robust enough against variation

• the combination of different methods to increase context reduces WER on a task of spontaneous telephone speech by 12.8% relative

• the problem is far from being solved: in the improved system, the
  ★ WER for children is 93% above average
  ★ WER for non-native speakers is 86% above average

• future work is to combine adaptive training methods with feature extraction methods that take into account context

• study the effects of variation on the features using as many different speech corpora as possible