JOINT OPTIMIZATION OF EVENT DETECTORS AND EVIDENCE MERGER FOR CONTINUOUS PHONE RECOGNITION

Marco Siniscalchi, Øystein Birkenes, Magne H. Johnsen, Torbjørn Svendsen

Norwegian University of Science and Technology
{marco77,birkenes,mhj,torbjorn}@iet.ntnu.no
Outline

✓ Sirkus Project at a Glance
✓ Attribute Detectors & Event Merger
✓ Joint Training of Detectors and Merger
✓ N-Best List Rescoring
✓ Experimental Results
Outline

✓ Sirkus Project at a Glance
✓ Attribute Detectors & Event Merger
✓ Joint Training of Detectors and Merger
✓ N-Best List Rescoring
✓ Experimental Results
Sirkus Project

A detection-based paradigm inspired by human speech recognition (HSR) system

Detect acoustic and auditory events

Merge events to form cognitive hypotheses

Validate the hypotheses using higher level knowledge

Today’s focus.
Speech Attributes

• Manner and Place of Articulation are events of interest because:
  – They are strictly related to human speech production,
  – They provide robustness to noise and cross-speaker variation,
  – They are portable across different languages,
  – They explicitly model linguistic information, which makes it easier to deal with non-native and hyper-articulated speech.
  – They can be combined with MFCCs
    – very useful for robust ASR applications.
Outline

✓ Sirkus Project at a Glance
✓ **Attribute Detectors & Event Merger**
✓ Joint Training of Detectors and Merger
✓ N-Best List Rescoring
✓ Experimental Results
A different parametrization for each detector is possible. Either frame-based or segment-based detectors can be used. Each output informs about the probability of the presence of an attribute in the given data.
Segment-Based Speech Detectors

Detectors implemented using competing HMM-based architecture.

Target HMM trained on the “feature present” segments

Non-target HMM trained on the “feature absent” segments

A collection of 15 dedicated detectors for manner and place of articulation is used.
How to Build Non-Target HMM Model

Several way to define the competing class.

Most competitive imposter implies the use of LLR.

All the other classes (universal background model) implies the use of LLR.

Cohort set implies the use of generalized LLR (GLLR).
Log-likelihood ratio (LLR) is used as measure to score input speech segments.

\[ LLR(x) = \log \frac{P(x | \lambda)}{P(x | \lambda^a)} \]
Detectors with Complex Competing Class

GLLR is distance between two collections of models, simple vs. composite hypotheses.

<table>
<thead>
<tr>
<th>Target</th>
<th>Cohort of five</th>
</tr>
</thead>
<tbody>
<tr>
<td>“w”</td>
<td>“l”, “el”, “ao”, “ow”, “uw”</td>
</tr>
<tr>
<td>“ah”</td>
<td>“aa”, “ax”, “ow”, “eh”, “aw”</td>
</tr>
<tr>
<td>“n”</td>
<td>“en”, “m”, “ng”, “ix”, “l”</td>
</tr>
</tbody>
</table>

(“w-ah+n” is the phoneme sequence for the word “one”)

\[ \ell(X \mid \lambda) \] is the likelihood of \( X \) given \( \lambda \)

\[
\text{GLLR}(X \mid \lambda_q, \Lambda_q) = \log[\ell(X \mid \lambda_q)] - \log[f(\ell(X \mid \Lambda_q))]
\]

\[
f(\ell(X \mid \Lambda_q)) = \left\{ |C_q|^{-1} \sum_r \exp\{[\eta \log \ell(X \mid \lambda_r)]\}\right\}^{1/\eta}
\]

(\( |C_q| \) is the size of the cohort set \( C_q \) of the claimed target \( q \))
Event Merger

Event Merger generates high level evidence.

Combines temporal events, e.g. combining attributes into phones.

Combines spatial events, e.g. combining vowel and nasal into nasalized vowels.

Probabilistic attribute lattice

Probabilistic phone/word lattice

Merger

A linear function is used to implement the merger.
Outline

 ✓ Sirkus Project at a Glance
 ✓ Attribute Detectors & Event Merger
 ✓ **Joint Training of Detectors and Merger**
 ✓ N-Best List Rescoring
 ✓ Experimental Results
Penalized logistic regression machine (PLRM) is used because it provides a framework that:

– Is statistically well founded,

– Makes easy definition and optimization of a criterion function that is jointly dependent upon the detector parameters and the merger parameters

– Allows an easy definition of joint criterion function

– Supports discriminative training

– Allows generation of posterior probabilities either at a phone or a word level.
The PLRM with Detector-Based Regressors

PLRM estimates the conditional probability of a phone (word) label $y$ given a speech segment $x$ as:

$$
\hat{p}_k = \hat{p}(y = k \mid x, W, \Lambda) = \frac{\exp w_k^T \phi(x; \Lambda)}{\sum_{i=1}^{K} \exp w_i^T \phi(x; \Lambda)}
$$

The regressor (detector) implementation is based on competing HMMs.
• Universal background model is used for the non-target HMM.
• LLR is used as goodness-of-fit between the input and the output of each detector.
Penalized Logistic Regression Objective Function

A joint optimization function in terms of the parameters of the detectors and merger can be defined as:

\[
P_\delta^{\log} (W; \Lambda; D) = \sum_{l=1}^{L} \log \hat{p}_{y(l)} + \frac{\delta}{2} \text{trace} \Gamma W \Sigma W^T
\]

**Logistic regression likelihood**

*Penalty term***

\( L \): the number of training segments

\( D = (x(l), y(l))_{l=1...L} \): the training set

\( \delta \): the penalty factor

\( \Sigma = (1/L) \Phi^T \Phi \),

\( \Phi \) is an \((M+1) \times L\) matrix with columns \( \Phi(x(l), \Lambda) \).

\( \Gamma \): \( K \times K \) diagonal matrix whose diagonal element accounts for the number of training segments in class \( k \).
Coordinate Descent for PLRM Training

\[
\min( P_\delta^{\log} (W ; \lambda ; D)) = \min( \sum_{l=1}^{L} \log \hat{p}_{y(l)} + \frac{\delta}{2} \text{trace} \; \Gamma W \Sigma W^T )
\]

is equivalent to

\[
(W^*, \lambda^*) = \arg\min_{(W, \lambda)} P_\delta^{\log} (W, \lambda; D)
\]

Coordinate descent approach with coordinate \( W \) and \( \lambda \)

Inizialize \( \lambda_0 \) by MLE

\[
W_0 = \arg\min_{(W, \lambda)} P_\delta^{\log} (W, \lambda_0; D)
\]

\[
\lambda_{i+1} = \arg\min_{\lambda} P_\delta^{\log} (W_i, \lambda; D)
\]

\[
W_{i+1} = \arg\min_{W} P_\delta^{\log} (W, \lambda_{i+1}; D)
\]
Preliminary Experiment on Phone Classification

- The PLRM approach is validated on a phone classification problem.
- Experiments on the TIMIT corpus.
  - Only segments having a minimum duration of 30ms are considered.
  - The core set was used during the testing phase.

<table>
<thead>
<tr>
<th>System</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLRM/Attribute-based Regressors</td>
<td>75.5%</td>
</tr>
<tr>
<td>HMM-based Phone (MMI)</td>
<td>74.7%</td>
</tr>
</tbody>
</table>

More attributes can be defined, and more sophisticated models can be used (cohort)
Outline

- Sirkus Project at a Glance
- Attribute Detectors & Event Merger
- Joint Training of Detectors and Merger
- **N-Best List Rescoring**
- Experimental Results
Rescoring as a means to provide additional information over what is used in conventional log-likelihood based decoding schemes.

\[
\hat{p}_{\hat{s}} = \left( \prod_{l=1}^{L_{\hat{s}}} p_{\hat{y}(l)} \right)^{1/L_{\hat{s}}}
\]

- Rescoring
- Geometric Mean
- Rescored N-Best
- Length of the sentence
“Garbage” Segments in N-Best List

• A speech segment in an N-best list may not correspond to a complete phone (or word), e.g:
  – The segment contains only a part of a phone (or word), or
  – The segment covers several phones (or words)

• Define and train a “garbage” class
  – Forced alignment given known transcription
  – N-best recognition
  – N-best segments that have $> \epsilon$ frames that differ from any FA segment are defined as “garbage”

• The training segments are grouped as:

$$\mathcal{D} = \mathcal{D}_{\text{correct}} \cup \mathcal{D}_{\text{garbage}}$$
Outline

✓ Sirkus Project at a Glance
✓ Attribute Detectors & Event Merger
✓ Joint Training of Detectors and Merger
✓ N-Best List Rescoring
✓ Experimental Results
Experimental Setup (1)

- TIMIT corpus is used for training and evaluation purposes, but SA sentences are not used.
- The NIST 24-speaker set is used during evaluation (192 sentences).
- Two baseline systems are built:
  - HTK is used to design a 3-state HMM/GMM phone recognizer with 16 mixture components per state.
  - BUT is used to design a hybrid 3-state HMM/ANN phone recognizer.
Experimental Setup (2)

- Garbage segments are generated from 20-best list for the 3696 training sentences.
- PLRM training is stopped after 6 coordinate descent iterations.
- $\delta$ is set equal to $10^{-1}$.
- Performance is reported in terms of relative improvement ($\text{Rel. Impr.}$), which is computed as

$$\text{Rel. Impr.} = \frac{\text{Acc}_{PLRM} - \text{Acc}_{Baseline}}{\text{Acc}_{NBest} - \text{Acc}_{Baseline}}$$

Achieved improvement

Achievable improvement
Results (1)

- HMM/GMM system is used as baseline.

<table>
<thead>
<tr>
<th>System</th>
<th>baseline</th>
<th>20-best</th>
<th>50-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>65.61%</td>
<td>70.30%</td>
<td>71.83%</td>
</tr>
</tbody>
</table>

- Joint optimization allows a relative improvement of 26.5%.

<table>
<thead>
<tr>
<th>Rescoring</th>
<th>20-best</th>
<th>50-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disjoint training</td>
<td>65.56%</td>
<td>65.42%</td>
</tr>
<tr>
<td>Joint training</td>
<td>66.85%</td>
<td>67.27%</td>
</tr>
</tbody>
</table>
- HMM/ANN system is used as **baseline**.

<table>
<thead>
<tr>
<th>Rescoring</th>
<th>20-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>74.42%</td>
</tr>
<tr>
<td>20-best</td>
<td>79.14%</td>
</tr>
<tr>
<td>PLRM</td>
<td>75.21%</td>
</tr>
<tr>
<td>Rel. Impr.</td>
<td>16.53%</td>
</tr>
</tbody>
</table>
Conclusions

• The detector-based approach is a highly flexible architecture.
• Attributes allow no direct modeling of phones.
• Joint training leads to high phone classification accuracy.
• Preliminary studies have shown that PLRM improves over both generative and discriminative models.
• Better selection of the PLRM parameters may lead to further improvements.
THANK YOU!

QUESTIONS?
PRELIMINARY RESULTS

- Experiments have been carried out on the TIMIT core set.
- No extensive parameter tuning has been performed.
- Garbage segments have been generated using 100-best list for the training sentences.
- 20-best list is used during rescoring.
- Only 15 attribute detectors have been used as regressors.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$10^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUT-Baseline</td>
<td>74.42</td>
</tr>
<tr>
<td>PLRM</td>
<td>75.21</td>
</tr>
<tr>
<td>Upper-Bound</td>
<td>79.14</td>
</tr>
<tr>
<td>Rel. Impr.</td>
<td>16.73</td>
</tr>
</tbody>
</table>