Reasons for TNO to join RT05s/spkr

- Running Broadcast News speaker segmentation/clustering speech recognition system for Dutch since 2001†
  - segmentation necessary for
    - on-line processing
    - feature stream time reversal in Abbot acoustic NN
    - low latency
    - poor clustering
- Active in NIST speaker recognition evaluations since 2003
- Takes part in AMI EU meeting project
  - scenarios, data collection, data processing, interpretation, presentation
    - speaker segmentation/clustering
- Problem definition
- Evaluation measures

†http://speech.tm.tno.nl/radio1/
Speaker Diarization Error rate (SDE)

- correct speaker time
- missed speaker time
- misclassified time
- false alarm speaker time

\[ \text{spoken time} \]
\[ \text{speaker error time} \]
Speech Activity Detection a necessity

- Speaker diarization error rate
  - Error speaker time / spoken time

- Without speech activity detection:
  - All non-speech time is *false alarm* speaker error time
  - Total time $T$, spoken time $T_s$

\[
SDE > \frac{T - T_s}{T_s}
\]

- typical meeting scenario  
  
\[
\frac{T_s}{T} \approx \frac{2}{3} \implies SDE > 50 \%
\]

- SAD important in RT05s speaker diarization
  - ICSI offered SAD output to us
  - contrastive SPKR condition
SAD approaches

1) Energy based
   - e.g., all frames with energy > 20 dB under meeting maximum
   - works fairly well for telephone speech, speaker recognition
   - doesn't work with distant microphone
   - SAD error ≃ 50%

2) Two-phone speech recognition system
   - speech + non-speech 3-state LtoR phone models
   - Sonic decoder, 2-phone grammar
     - no output

3) Two-state Viterbi GMM decoder “ptsamiditw”
   - 16 mixtures/model
   - calculate maximum likelihood state sequence
   - apply some smoothing
   - seems to work
SAD results ptsamiditw

- GMM training, 12 PLP+energy+delta
  - 5 “train” AMI meetings from dev test
  - non/speech labels from SPKR reference files
    - thanks Xavier Anguera, ICSI
- decoder parameter tuning
  - 5 “test” AMI meetings from dev test
  - parameters
    - prior odds non/speech 0.01
    - transition probability ratio $10^{-5}$
- Results
  - SAD error rate
    - AMI dev test set 10.3 %
    - RT04s – CMU 2.8 %
    - RT05s 5.0 %
Speaker Segmentation

- Uses output from SAD, 12 PLP+energy
- Based on Bayesian Information Criterion, Chen&Gopalakrishnan†

\[
\Delta \text{BIC} = \frac{1}{2} \left( N_x \log |\Sigma| - N_A \log |\Sigma_A| - N_B \log |\Sigma_B| - \lambda N_M \log N_x \right)
\]

- \( N_x = N_A + N_B \) number of frames considered in current “window”

- store aggregated “sufficient statistics” for covariances

†Proc. DARPA broadcast news transcription and understanding, 1998
Speaker clustering

- Uses output from speaker segmentation
- Agglomerative clustering
- Uses “Gish distance measure” for finding closest segments

\[ G(c_i, c_j) = \frac{1}{2} \left( (N_i + N_j) \log |\Sigma_m| - N_i \log |\Sigma_i| - N_j \log |\Sigma_j| \right) \]

- Condition for merging clusters based on BIC

\[ \frac{1}{2} \lambda_{N_M} \log N_x - G(c_i, c_j) > 0 \]

- \( N_x \) is total number of frames in entire meeting
- Inefficient for large number of initial segments
  - but preferred over “online” version of BN system
- Tuning parameters
  - AMI “test” split development test data
  - \( \lambda_{\text{seg}} = 1.5 \quad \lambda_{\text{clust}} = 14 \)
NIST RT05s speaker diarization results

- “Multiple distant microphones” = single distant mic
- no overlap
- SDE, in %

<table>
<thead>
<tr>
<th>Test set</th>
<th>TNO</th>
<th>ICSI</th>
<th>perfect</th>
<th>optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI dev</td>
<td>35.7</td>
<td>45.9</td>
<td>45.3</td>
<td>?</td>
</tr>
<tr>
<td>RT04s – CMU</td>
<td>35.4</td>
<td>31.9</td>
<td>25.6</td>
<td></td>
</tr>
<tr>
<td>RT05s</td>
<td>35.1</td>
<td>37.1</td>
<td>32.3</td>
<td>19.0</td>
</tr>
</tbody>
</table>

- RT05s speaker misses, false alarms
  - misses: 13/53 = 24.5% speakers, 0.4% speaker time
  - false alarms: 5/53 = 9.4% speakers, 6.6% speaker time
Discussion / conclusions

- SDE Evaluation measure
  - harsh on $T_{FA}$ because $T - T_{FA}$ in denominator
  - weights long duration speakers more
    - advantageous to ignore short duration speakers
    - high $\lambda_{clust}$
- BIC segmentation / clustering
  - nice idea based on first principles
  - still tunable parameters $\lambda$
  - why full covariance single mixture GMMs?
    - cancellation of exponent in likelihood calculation

\[
\log \prod_i N(x_i, \mu, \Sigma) = - \sum_i \log(2\pi)^{d/2} \sqrt{\det\Sigma} - \frac{1}{2} \sum_i (x - \mu)^T \Sigma^{-1}(x - \mu)
\]

- how about diagonal covariance, multiple mixtures?
No time / plans for next evaluation

- Use decoder for clustering process
  - use diagonal covariance GMM for speaker model
  - include overlap between speakers in network

- Use multiple distant microphone data
  - SAD: results from ICSI
  - SPKR: RT05s results not hopeful

- Investigate “absolute speaker ID”
  - “speaker spotting”
  - speaker tracking
  - speaker priors and evaluation measure
    - speaker speaking time entropy?†

†Jin et al., Proc NIST RT04s, ICASSP, 2004