AMIDA RT07s speaker diarization system

Matej Konečný and David van Leeuwen
Changes w.r.t. RT06s

- Project, name
  - EU AMI ended, EU AMIDA started
    - DA: Distant Access (not like MDM:)
- Personel
  - Marijn Huijbregts went from AMI to ICSI
  - Matej Konečný AMI/DA trainee from Brno with TNO
- Algorithm
  - Use MDM beamforming
    - signal enhancement
    - delay parameters
  - Cross Likelihood Ratio-based clustering (SID)
    - ‘no more tunable parameters’
    - Minimum duration Viterbi-decoder
- Tasks
  - No lecture room, no more SAD :(
  - Segmentation/clustering for STT, SASST
Overview

- Differences
- This overview
- Overview SPKR approach
- SAD experiments
- Overlap detection experiments
- Conclusions
AMIDA System design (Matej)

Data enhancement

Multiple microphone channels

Delay & Sum

Combined acoustic channel

Wiener filter

Feature extraction

Delay features

Acoustic features

Initialization

Speech Activity Detection

BIC segm. & clustering

Initial clustering

Viterbi & EM algorithm

MAP adaptation

CLR calculation

Merge clusters

Main stage

CLR > 0.0
RT07s system: mix of choices

- Speech activity detection
  - Wiener filter
  - Initial segmentation
  - Re-segmentation / clustering

- Speaker/cluster modeling
  - Segmentation
    - Gaussian Mixture Models, #Gaussians(size data)
  - Cluster criterion
    - UBM-GMM, Cross-Likelihood Ratio
System design: front end processing

- Delay and sum beamforming
  - Use Xavie’s BeamformIt 2.0
  - use only 32 ms window and 16 ms stepsize
    - different from 500 ms / 250 ms default
    - aligned with PLP feature extraction
  - Use Wiener filtering noise reduction
    - after beamforming
  - Qualcomm-ICSI-OGI toolkit
  - SAD from toolkit

- Use SAD trained on
  - 10 AMI meetings from RT05s development, SDM
  - not beamformed/filtered
System design: features and modeling

- 13 PLP features (no derivatives)
  - ICSI / Dan Ellis’ *rasta* tool
- \( N-1 \) delay parameters from delay&sum
  - \( N \) microphones in MDM
- Speaker/cluster modeled by Gaussian Mixture Model
  - 1 Gaussian for delay parameters
  - 1–64 Gaussians for PLP features
    - Cluster complexity ratio ~ 300
    - 4.8 sec speech / Gaussian
- Initialization of GMMs
  - doubling \( N_G \) until power of 2 below desired \( N_G \)
  - Iteratively increasing \( N_G \) by one
Segmentation

• Initialization
  • Generate initial segments using BIC segmenter / clusterer
    • $\lambda_{BIC} = 1$ for both
    • many short segments
    • many small clusters
  
• Use segmentation for training initial GMMs for diarization

• Viterbi re-segmentation (5x)
  • decode
  • keep track of $N_G$ for each cluster dependent on amount of data
    • 4.8 sec / Gaussian
  • grow $N_G$ by splitting
  • reduce $N_G$ by retraining GMM from scratch
Clustering

- Build 64 Gaussian UBM from entire meeting (once)
- MAP adapt UBM to data found by segmentation
- compute cross likelihood ratio for each pair of clusters
  \[ R_{ij} = \frac{1}{n_i} \log \frac{p(x_i | \lambda_j)}{p(x_i | \lambda_{UBM})} + \frac{1}{n_j} \log \frac{p(x_j | \lambda_i)}{p(x_j | \lambda_{UBM})} \]
- Merge clusters \( i \) and \( j \) for which
  - \( R_{ij} \) is largest and
  - positive
- Stop if maximum \( R_{ij} < 0 \)
Progress, effect of delay parameters

<table>
<thead>
<tr>
<th>System</th>
<th>DER RT05s (overlap)</th>
<th>DER RT06s (overlap)</th>
<th>DER RT07s (overlap)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI RT06s</td>
<td>21.7%</td>
<td>32.4%</td>
<td>26.2%</td>
</tr>
<tr>
<td>AMIDA RT07s primary</td>
<td>16.3%</td>
<td>18.1%</td>
<td>22.0%</td>
</tr>
<tr>
<td>AMIDA RT07s no delay params</td>
<td>20.5%</td>
<td>24.3%</td>
<td></td>
</tr>
</tbody>
</table>

- System has become slightly more robust
- But there still is high variability along dataset
- Delay parameters seem to help quite a bit
Another SAD story

• Good history in Speech Activity Detection performance
  • using 10 AMI meetings for modeling non/speech
  • SDM

• This year using Forced Aligned reference non/speech
• Also using Beamforming/MDM

• Two sets of non/speech models
  • (1) original SDM AMI RT05s-dev
  • (2) new RT05/RT06 FA MDM beamformed

• Best results (mixsad)
  • using (1) for BIC segmentation/clustering
  • using (2) for final frame selection
Results 2006/2007, effect of Speech Activity Detection

<table>
<thead>
<tr>
<th>BIC seg/clust SAD</th>
<th>Final SAD</th>
<th>DER RT06s (overlap)</th>
<th>DER RT07s (overlap)</th>
<th>DER RT07s (no overlap)</th>
<th>SAD err</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>AMI</td>
<td>18.1%</td>
<td>22.0%</td>
<td>18.9%</td>
<td>6.7%</td>
</tr>
<tr>
<td>AMI</td>
<td>RT forced alignment</td>
<td>20.1%</td>
<td>17.0%</td>
<td>13.4%</td>
<td>2.9%</td>
</tr>
<tr>
<td>RT forced alignment</td>
<td>RT forced alignment</td>
<td>18.6%</td>
<td>15.3%</td>
<td>2.9%</td>
<td></td>
</tr>
</tbody>
</table>

- DER very dependent on SAD
- Still no consistent behaviour between RT years
- Still a lot depends on initialization of GMMs
Overlapping speech approach

- Two steps:
  - overlap detection
  - overlapping speaker attribution

- Cheating experiment:
  - perfect overlap detection
  - assign most talkative speaker as 2nd speaker
  - about 2% reduction in DER

- Overlap detection
  - BeamformIt: 6.65% FA @ 85.7% miss
    - $d' = 0.2$, or EER = 46%
    - not good enough detection
  - Training GMMs with/out overlapping speech, decode
  - Building ‘overlapping’ GMMs from ‘single’ clusters

Figure 2: A DET plot, showing the trade-off between false alarm and miss probabilities. The operating point of the decisions made, the co-ordinates correspond to the surface of the grey areas in Figure 1.

Chapter 2
System description

Figure zvy: General concept of speaker diarization system

Most of the approaches from... and BICu-based initialization take partv Figure zvz depicts the system design. The steps can be described as...
Conclusions

• Front-end processing finally pays off
  • SNR improvement
    • delay&sum
    • Wiener filter
  • Modeling of Delay parameters helps

• Initialization of GMMs seems to be important
  • used deterministic estimation this year

• Hardly any ‘tunable parameters’
  • Cluster complexity ratio

• SAD still very important

• Overlapping speech still is a challenge