The AMI RT09s Speaker Diarization System

Marijn Huijbregts, David van Leeuwen, Thomas Hain
Outline

- System overview
- Post evaluation analysis
- Current and future work
The AMI RT09s speaker diarization system

System overview

Pre-processing MDM:

Recording
- Microphone 1
- Microphone 2
- Microphone N

Noise reduction
Wiener-filtering, Aurora toolkit

Beam forming
(BeamformIt 2.0)

Single channel recording
Channel delay features
Pre-processing SDM:

- Recording
- Microphone 1
- Noise reduction
  - Wiener-filtering, Aurora toolkit
- Single channel recording

The AMI RT09s speaker diarization system
The AMI RT09s speaker diarization system

System overview

The MDM system:

- Speech/non-speech (developed for RT07s at ICSI)
- Speech/non-speech segmentation
- Channel delay features
- Speaker segmentation
- Agglomerative clustering (improved RT06s system)

Single channel recording

Speaker A
Speaker B
Speaker C
Sp. A
Speaker B
Feature extraction (for the clustering component):

- 19 MFCC features (no delta's, no energy)
- Delay features (dependent on #microphones)
- Mean and variance normalization
- Variance flooring (0.01)
Speaker modeling:

- Number of initial clusters depends on amount of data: #speech-vectors / 4000
- 5 Gaussians for MFCC, 1 for delay feature stream
- The two feature streams are normalized in two ways:
  - Stream weights are determined on basis of inverse entropy of BIC scores (only first 2 iterations, weights are fixed after that)
  - The average BIC score of the delay stream is normalized to the average BIC score of the MFCC stream, comparable to (cepstrum) mean normalization. (only first 2 iterations)
The AMI RT09s speaker diarization system

System overview

The SDM system:

Speech/non-speech (developed for RT07s at ICSI) → Speech/non-speech segmentation

Single channel recording → Agglomerative clustering (improved RT06s system) → Speaker segmentation

The AMI RT09s speaker diarization system
Meeting room assignment (MDM):

- The EDI main and remote meeting rooms are processed separately.
- For each final cluster in both rooms, it is investigated in which room the speech occurs first. All speech occurring earlier in another room is removed from the room-clustering.
- The main and remote speaker transcriptions are then merged.
A lot of submissions...

We have submitted a lot of systems, mainly for use in our Interspeech paper

- **mdm_c-BeamFormedDelay**: Baseline MDM system (used for ASR)
- **mdm_c-BeamFormedBaseline**: MDM, without delay features
- **mdm_c-onestreamnoroom**: MDM, no delay features, only main meeting room
- **sdm_c-sad1**: Baseline SDM system

The other submissions are part of current research on overlap detection and system combination (more on that later)
Evaluation results:

- Single Distant Microphone: 29.0 %DER
- Multiple Distant Microphone: 21.5 %DER

This improvement must be due to:

- Beam forming
- Delay features
- EDI room selection
Evaluation results:

- **Multiple Distant Microphone:** 21.5 %DER
  (mdm_c-BeamFormedDelay)

- **MDM, only main meeting room:** 23.2 %DER
  (post-eval experiment)

- **MDM, no delay features:** 28.8 %DER
  (mdm_c-BeamFormedBaseline)

- **Single Distant Microphone:** 29.0 %DER
  (sdm_c-sad1)

The delay features seem to help, but beam forming itself does not?
## Post evaluation analysis

<table>
<thead>
<tr>
<th></th>
<th>One room</th>
<th>no delay</th>
<th>SDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDI_20071128-1000</td>
<td>18.40</td>
<td>19.03</td>
<td>19.53</td>
</tr>
<tr>
<td>EDI_20071128-1500</td>
<td>16.75</td>
<td>18.16</td>
<td>41.98</td>
</tr>
<tr>
<td>IDI_20090128-1600</td>
<td>21.54</td>
<td>22.86</td>
<td>23.04</td>
</tr>
<tr>
<td>IDI_20090129-1000</td>
<td>21.58</td>
<td>21.85</td>
<td>19.21</td>
</tr>
<tr>
<td>NIST_20080201-1405</td>
<td>60.45</td>
<td>61.46</td>
<td>56.19</td>
</tr>
<tr>
<td>NIST_20080227-1501</td>
<td>24.73</td>
<td>35.57</td>
<td>35.88</td>
</tr>
<tr>
<td>NIST_20080307-0955</td>
<td>11.32</td>
<td>38.26</td>
<td>21.34</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>23.17</strong></td>
<td><strong>28.75</strong></td>
<td><strong>28.80</strong></td>
</tr>
</tbody>
</table>
### Post evaluation analysis

<table>
<thead>
<tr>
<th>Speaker A</th>
<th>Speaker B</th>
<th>Speaker C</th>
<th>Sp. A</th>
<th>Speaker B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDI_20071128-1000</td>
<td>18.40</td>
<td>19.03</td>
<td>19.53</td>
<td></td>
</tr>
<tr>
<td>EDI_20071128-1500</td>
<td>16.75</td>
<td>18.16</td>
<td>41.98</td>
<td></td>
</tr>
<tr>
<td>IDI_20090128-1600</td>
<td>21.54</td>
<td>22.86</td>
<td>23.04</td>
<td></td>
</tr>
<tr>
<td>IDI_20090129-1000</td>
<td>21.58</td>
<td>21.85</td>
<td>19.21</td>
<td></td>
</tr>
<tr>
<td>NIST_20080201-1405</td>
<td>60.45</td>
<td>61.46</td>
<td>56.19</td>
<td></td>
</tr>
<tr>
<td>NIST_20080227-1501</td>
<td><strong>24.73</strong></td>
<td><strong>35.57</strong></td>
<td>35.88</td>
<td></td>
</tr>
<tr>
<td>NIST_20080307-0955</td>
<td>11.32</td>
<td>38.26</td>
<td>21.34</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>23.17</td>
<td>28.75</td>
<td>28.80</td>
<td></td>
</tr>
</tbody>
</table>

Delay features don't help much either, except in two NIST meetings.

The AMI RT09s speaker diarization system
Beam forming actually *decreases* performance for a number of meetings.

<table>
<thead>
<tr>
<th></th>
<th>Speaker A</th>
<th>Speaker B</th>
<th>Speaker C</th>
<th>Sp. A</th>
<th>Speaker B</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDI_20071128-1000</td>
<td>18.40</td>
<td>19.03</td>
<td>19.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDI_20071128-1500</td>
<td>16.75</td>
<td>18.16</td>
<td>41.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDI_20090128-1600</td>
<td>21.54</td>
<td>22.86</td>
<td>23.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDI_20090129-1000</td>
<td>21.58</td>
<td>21.85</td>
<td>19.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST_20080201-1405</td>
<td>60.45</td>
<td>61.46</td>
<td>56.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST_20080227-1501</td>
<td>24.73</td>
<td>35.57</td>
<td>35.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST_20080307-0955</td>
<td>11.32</td>
<td>38.26</td>
<td>21.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>23.17</td>
<td>28.75</td>
<td>28.80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The AMI RT09s speaker diarization system
The improvements for speech/non-speech are limited as well:

- SDM (evaluation, development): 5.12% 4.24%
- MDM (evaluation, development): 4.45% 3.35%
- Relative improvement: 13% 21%

We need to analyze this further to understand why beam forming is not helping us as much as in earlier evaluations.
We have done some initial research on:
- overlap detection
- combining system outputs

(Hopefully) this work will be published at Interspeech 2009

The improvement due to overlap detection was limited (0.2%DER) and the system combination did not help (degradation of 0.79%DER)

We did not perform an analysis yet on why these two techniques did not work.
Preliminary on-line approach

• On-line restrictions come in at many stages:
  – SAD (possible with small delay)
  – Wiener filtering (possible with prior noise estimate)
  – Beamforming (possible at level of analysis window)
  – Segmentation --- not really time for that, segmentation means “limited history for current output decision”
  – Clustering --- on-line variant not necessarily much worse than agglomerative
In future work we will focus on:

- Improving our overlap detection and system combination methods
- Apply other feature types as second feature stream (especially for SDM)
- Creating a system that is able to process longer meetings (our current system is not scalable)
### Conclusions

- We've seen again
  - Eval set more “difficult” than dev set
  - despite large and heterogeneous dev set (27 meetings)
  - despite principle of don't-use-any-prior-information
  - apparently still lots of “system-tuning” specific for devset
    - algorithm choice
    - algorithm hyper-parameters
  - perhaps due to “out-of-domain”
    - meeting duration, meeting room layout
Questions?

(or email Marijn for questions: marijn.huijbregts@utwente.nl)