A DOA BASED SPEAKER DIARIZATION SYSTEM FOR REAL MEETINGS

Araki S., Fujimoto M., Ishizuka K., Sawada H. and Makino S.
NTT Communication Science Laboratories

Listen!
25\textsuperscript{th} January 2010
Presented by Erich Zwyssig
CSTR/ EADS IW UK Ltd.
Contents

- Abstract
- Introduction
- Method
- Setup
- Analysis
- Results
- Conclusions
- References/Demo
Abstract

- This paper presents a speaker diarisation system that estimates who spoke when in a meeting. Our proposed system is realized by using a noise robust voice activity detector (VAD), a direction of arrival (DOA) estimator, and a DOA classifier. Our previous system utilized the generalized cross correlation method with the phase transform (GCC-PHAT) approach for the DOA estimation. Because the GCC-PHAT can estimate just one DOA per frame, it was difficult to handle speaker overlaps. This paper tries to deal with this issue by employing a DOA at each time-frequency slot (TFDOA), and reports how it improves diarisation performance for real meetings / conversations recorded in a room with a reverberation time of 350 ms.
Introduction

- Speaker Diarisation (Who spoke when)
- Useful for speech recognition in:
  - Minute taking
  - Speech enhancement
  - ...
- RT (real-time) processing
- GCC-PHAT DOA estimator implementation only detects one speaker per frame
Assumption

- Conversational speech
- Overlapping speech
- Speaker do not change seats during one meeting/conversation
- SNR sufficiently high to safely detect speech looking at energy only
Method

- Previous
  - missed speakers (when overlapping)
  - directional noise is detected as speech (leading to missed low-energy speech)
  - difficult to set VAD threshold

- New
  - improved DOA scheme → TFDOA
  - DOA amplitude weighting
  - probabilistic VAD
Method

Fig. 1. Block diagram of previous method (framed by dashed line) and proposed method (Method I).
System Evaluation

- **Setup**
  - 3 microphones
  - 3-4 speakers
  - 5' meetings (4x) (large # of turns)

- **Evaluation measure**
  - DER = wrong estimation / entire speaker length (Diarisation Error Rate)

---

Table 1. Conversation recordings. Each recording lasted five minutes.

<table>
<thead>
<tr>
<th>Evaluation data ID</th>
<th>#Speaker</th>
<th>Overlap [%]</th>
<th>#Turn-taking</th>
<th>#Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>crossword puzzle 1</td>
<td>4</td>
<td>18.6</td>
<td>149</td>
<td>185</td>
</tr>
<tr>
<td>crossword puzzle 2</td>
<td>4</td>
<td>13.0</td>
<td>183</td>
<td>218</td>
</tr>
<tr>
<td>discussion</td>
<td>3</td>
<td>10.8</td>
<td>126</td>
<td>172</td>
</tr>
<tr>
<td>conversation</td>
<td>3</td>
<td>34.8</td>
<td>243</td>
<td>278</td>
</tr>
</tbody>
</table>
Setup
Results

Fig. 4. Diarization result examples. (a) ground truth label, (b) with previous method and (c) with Method I. Gray dots indicate noise frames.
Results

Fig. 5. Diarization result examples. (a) ground truth label, (b) with previous method, (c) with Method I and (d) with Method II. Gray dots show noise frames.
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>MST</th>
<th>FAT</th>
<th>SET</th>
<th>DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td>35.3</td>
<td>8.1</td>
<td>9.7</td>
<td>53.2 (41.1)</td>
</tr>
<tr>
<td>Method I</td>
<td>31.4</td>
<td>10.5</td>
<td>6.4</td>
<td>48.3 (37.0)</td>
</tr>
<tr>
<td>Method II</td>
<td>31.2</td>
<td>8.2</td>
<td>3.9</td>
<td>43.2 (28.5)</td>
</tr>
<tr>
<td>Method III</td>
<td>25.1</td>
<td>6.5</td>
<td>3.6</td>
<td>35.2 (21.9)</td>
</tr>
</tbody>
</table>

**Crossword puzzle 1**

<table>
<thead>
<tr>
<th>Method</th>
<th>MST</th>
<th>FAT</th>
<th>SET</th>
<th>DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td>28.4</td>
<td>12.7</td>
<td>8.9</td>
<td>50.1 (40.0)</td>
</tr>
<tr>
<td>Method I</td>
<td>24.2</td>
<td>17.8</td>
<td>5.2</td>
<td>47.2 (36.1)</td>
</tr>
<tr>
<td>Method II</td>
<td>23.9</td>
<td>15.1</td>
<td>3.1</td>
<td>42.2 (29.2)</td>
</tr>
<tr>
<td>Method III</td>
<td>19.6</td>
<td>13.4</td>
<td>2.6</td>
<td>35.5 (25.0)</td>
</tr>
</tbody>
</table>

**Crossword puzzle 2**

<table>
<thead>
<tr>
<th>Method</th>
<th>MST</th>
<th>FAT</th>
<th>SET</th>
<th>DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td>43.5</td>
<td>5.1</td>
<td>2.1</td>
<td>50.6 (35.9)</td>
</tr>
<tr>
<td>Method I</td>
<td>38.5</td>
<td>9.0</td>
<td>1.5</td>
<td>49.0 (31.9)</td>
</tr>
<tr>
<td>Method II</td>
<td>37.8</td>
<td>7.8</td>
<td>1.5</td>
<td>47.0 (32.6)</td>
</tr>
<tr>
<td>Method III</td>
<td>38.5</td>
<td>5.5</td>
<td>1.5</td>
<td>45.5 (29.9)</td>
</tr>
</tbody>
</table>

**Discussion**

<table>
<thead>
<tr>
<th>Method</th>
<th>MST</th>
<th>FAT</th>
<th>SET</th>
<th>DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td>38.2</td>
<td>7.5</td>
<td>8.0</td>
<td>53.7 (40.7)</td>
</tr>
<tr>
<td>Method I</td>
<td>30.0</td>
<td>13.1</td>
<td>7.6</td>
<td>50.7 (37.7)</td>
</tr>
<tr>
<td>Method II</td>
<td>30.4</td>
<td>12.3</td>
<td>6.2</td>
<td>48.9 (32.3)</td>
</tr>
<tr>
<td>Method III</td>
<td>33.2</td>
<td>9.5</td>
<td>6.1</td>
<td>48.8 (34.3)</td>
</tr>
</tbody>
</table>

**Conversation**

---

MST: missed  
FAT: false alarm  
SET: speaker error  
DER: error rate
Conclusions

- use DOA for diarisation (instead of GCC-PHAT)
- reduces missed speaker time → and therefore improves performance
- amplitude observation useful to disregard directional noise
- probabilistic VAD improves performance
Caveats

- fixed speakers
- tested on 5' meetings
- used only 4 meetings (?)
- Average utterance is 2-3 seconds (?)
- 2 manually determined thresholds
References

- Realtime Multimodal System for Conversation Scene Analysis (NTT RT System)

Other