Hidden Markov model diarisation with speaker location information

Jeremy Wong, Xiong Xiao, and Yifan Gong

Microsoft Speech and Language Group
Diarisation pipeline

Propose:
- Use speaker location information inside HMM clustering.
Prior work

• Use time-delay-of-arrival as observed variable in HMM [19, 20].
• Speaker location tracking using Kalman filters.

Sound source localisation

- Instantaneous speaker location is represented by Sound Source Localisation (SSL) vector, $s_t$.

$$s_{ti} = P(\theta_t = i|o_t)$$

- Estimate SSL vector from multi-channel audio using complex angular central Gaussian model [18].
- SSL explicitly represents where speaker is located.
- TDOA only implicitly captures speaker location information.

[18] T. Yoshioka et. al., “Advances in online audio-visual meeting transcription”, ASRU, 2019
HMM clustering

\[
p(D_{1:T}, S_{1:T}) \approx \sum_{q_{1:T}} \prod_{t=1}^{T} p^k(d_t | q_t)p^\gamma(s_t | q_t)P(q_t | q_{t-1})
\]

• Variables:
  ➢ \(d_t\): speaker embedding
  ➢ \(s_t\): SSL vector
  ➢ \(q_t\): HMM state, representing a speaker

• Assume that \(d_t\) and \(s_t\) are independent, given \(q_t\).
Speaker embedding emission

• Speaker embedding observation log-likelihood is cosine distance.

\[
\log p(d_t|q_t) = w_t d_t \cdot \mu_{q_t}
\]

• Variables:
  - \(\mu_{q_t}\): speaker embedding of HMM state \(q_t\) (HMM parameter)
  - \(w_t\): duration of segment \(t\)

• Equivalent likelihood is von-Mises Fisher density function.
Speaker location emission

• Proposed speaker location observation log-likelihood is KL-divergence.

\[ \log p(s_t | q_t) = w_t s_t \cdot \log \phi_{q_t} \]

• Variables:
  \( \phi_{q_t} \): average speaker location of HMM state \( q_t \) (HMM parameter)

• Equivalent likelihood is continuous categorical density function.
E-M estimation

• Auxilliary loss maximisation

$$\phi_i^{u+1} = \arg\max_{\phi_i} \sum_{t=1}^{T} P(q_t = i | D_{1:T}, S_{1:T}, \phi_i^u) \gamma w_t s_t \cdot \log \phi$$

s.t. $\phi_i \geq 0$ and $\sum_i \phi_i = 1$

• M-step update:

$$\phi_i^{u+1} = \frac{\sum_{t=1}^{T} P(q_t = i | D_{1:T}, S_{1:T}, \phi_i^u) w_t s_t}{\sum_j \sum_{t=1}^{T} P(q_t = i | D_{1:T}, S_{1:T}, \phi_i^u) w_t s_t}$$

• $\phi$ represents average location of speaker throughout meeting.
Diarisation steps

1. AHC clustering.
2. Initialise HMM parameters from AHC hypothesis.
4. Decode cluster sequence.
   \[ q^{*}_{1:T} = \arg \max_{q_{1:T}} \prod_{t=1}^{T} P(q_t | D_{1:T}, S_{1:T}) \]
5. Speaker tagging using the Hungarian algorithm.
Meeting transcription setup

• Multi-channel input.
• Post-ASR diarisation.

Data:
• dev: 51 meetings, 23 hours
• eval: 60 meetings, 35 hours
• Average of 7 participants per meeting

Speaker-attributed WER metric:
• Compute WER separately for each speaker.
• Average WERs over all speakers.
**Experiments**

<table>
<thead>
<tr>
<th>Use SSL</th>
<th>E-M fine-tune</th>
<th>Speaker-attributed WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>none</td>
<td>dev 22.75 eval 21.41</td>
</tr>
<tr>
<td>no</td>
<td>$\lambda, \eta$</td>
<td>dev 22.47 eval 21.15</td>
</tr>
</tbody>
</table>

**Uniformly initialize $\phi$**

| yes     | $\lambda, \eta, \phi$ | dev 21.62 eval 20.42        |

**Initialise $\phi$ from AHC hypothesis**

| yes     | $\lambda, \eta$    | dev 22.25 eval 20.55        |
| yes     | $\lambda, \eta, \phi$ | dev 21.61 eval 20.37    |

- SSL is complementary to speaker embeddings.
- E-M fine-tuning of HMM parameters is beneficial.
Experiments

- After E-M, $\phi$ resembles average speaker position throughout whole meeting.
- When speaker moves, $\phi$ becomes multi-modal.
Summary

Proposed:
• Incorporate speaker location into HMM diarisation.

Results:
• Speaker location is complementary to speaker embeddings.

Future work:
• Investigate speaker movement tracking in diarisation.