Microsoft

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.

Microsoft

Prior work

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

[19] J. Pardo et. al., "*Speaker diarization for multiple-distant-microphone meetings using several sources of information*", IEEE Transactions on Computers, vol. 56, no. 9, 2007
[20] D. Vijayasenan and F. Valente, "*Speaker diarization of meetings based on large TDOA feature vectors*", ICASSP, 2012

Sound source localisation

- Instantaneous speaker location is represented by Sound Source Localisation (SSL) vector, $\mathbf{s}_t$.

$$s_{ti} = P(\theta_t = i|\mathbf{o}_t)$$

- 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

$$p(\mathbf{D}_{1:T}, \mathbf{S}_{1:T}) \approx \sum_{\mathbf{q}_{1:T}} \prod_{t=1}^{T} p^{\kappa}(\mathbf{d}_t | q_t) p^{\gamma}(\mathbf{s}_t | q_t) P(q_t | q_{t-1})$$

- **Variables:**
  - $\mathbf{d}_t$: speaker embedding
  - $\mathbf{s}_t$: SSL vector
  - q_t: HMM state, representing a speaker

- Assume that $\mathbf{d}_t$ and $\mathbf{s}_t$ are independent, given q_t.

Speaker embedding emission

- Speaker embedding observation log-likelihood is cosine distance.

$$\log p(\mathbf{d}_t | q_t) = w_t \mathbf{d}_t \cdot \boldsymbol{\mu}_{q_t}$$

- **Variables:**
  - ➢ $\boldsymbol{\mu}_{q_t}$: 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.

$$\log p(\boldsymbol{s}_t | q_t) = w_t \boldsymbol{s}_t \cdot \log \boldsymbol{\phi}_{q_t}$$

- **Variables:**
  - $\boldsymbol{\phi}_{q_t}$: average speaker location of HMM state q_t (HMM parameter)

- Equivalent likelihood is continuous categorical density function.

E-M estimation

- Auxilliary loss maximisation

$$\phi_i^{u+1} = \operatorname*{argmax}_{\phi_i} \sum_{t=1}^{T} P(q_t = i \big| \mathbf{D}_{1:T}, \mathbf{S}_{1:T}, \phi_i^u) \gamma w_t \mathbf{s}_t \cdot \log \boldsymbol{\phi}$$

$$\text{s.t.} \quad \phi_i \geq 0 \quad \text{and} \quad \textstyle\sum_i \phi_i = 1$$

- M-step update:

$$\phi_i^{u+1} = \frac{\sum_{t=1}^{T} P(q_t = i \big| \mathbf{D}_{1:T}, \mathbf{S}_{1:T}, \phi_i^u) w_t \mathbf{s}_t}{\sum_j \sum_{t=1}^{T} P(q_t = i \big| \mathbf{D}_{1:T}, \mathbf{S}_{1:T}, \phi_i^u) w_t s_{tj}}$$

- $\boldsymbol{\phi}$ represents average location of speaker throughout meeting.

Microsoft

Diarisation steps

1. AHC clustering.
2. Initialise HMM parameters from AHC hypothesis.
3. Fine-tune HMM parameters on test meeting using E-M algorithm.
4. Decode cluster sequence.

$$\mathbf{q}_{1:T}^{*} = \underset{\mathbf{q}_{1:T}}{\operatorname{argmax}} \prod_{t=1}^{T} P(q_t | \mathbf{D}_{1:T}, \mathbf{S}_{1:T})$$

5. Speaker tagging using the Hungarian algorithm.

Microsoft

Meeting transcription setup

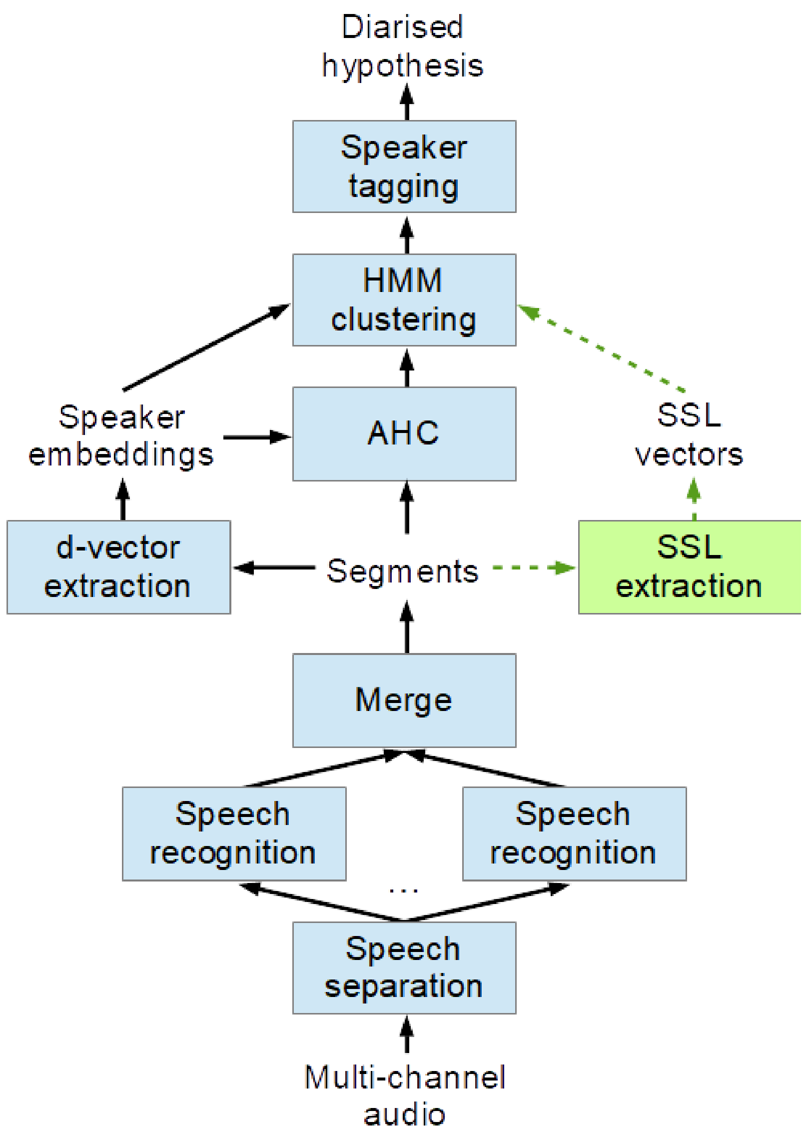
- Multi-channel input.
- Post-ASR diarisation.

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.



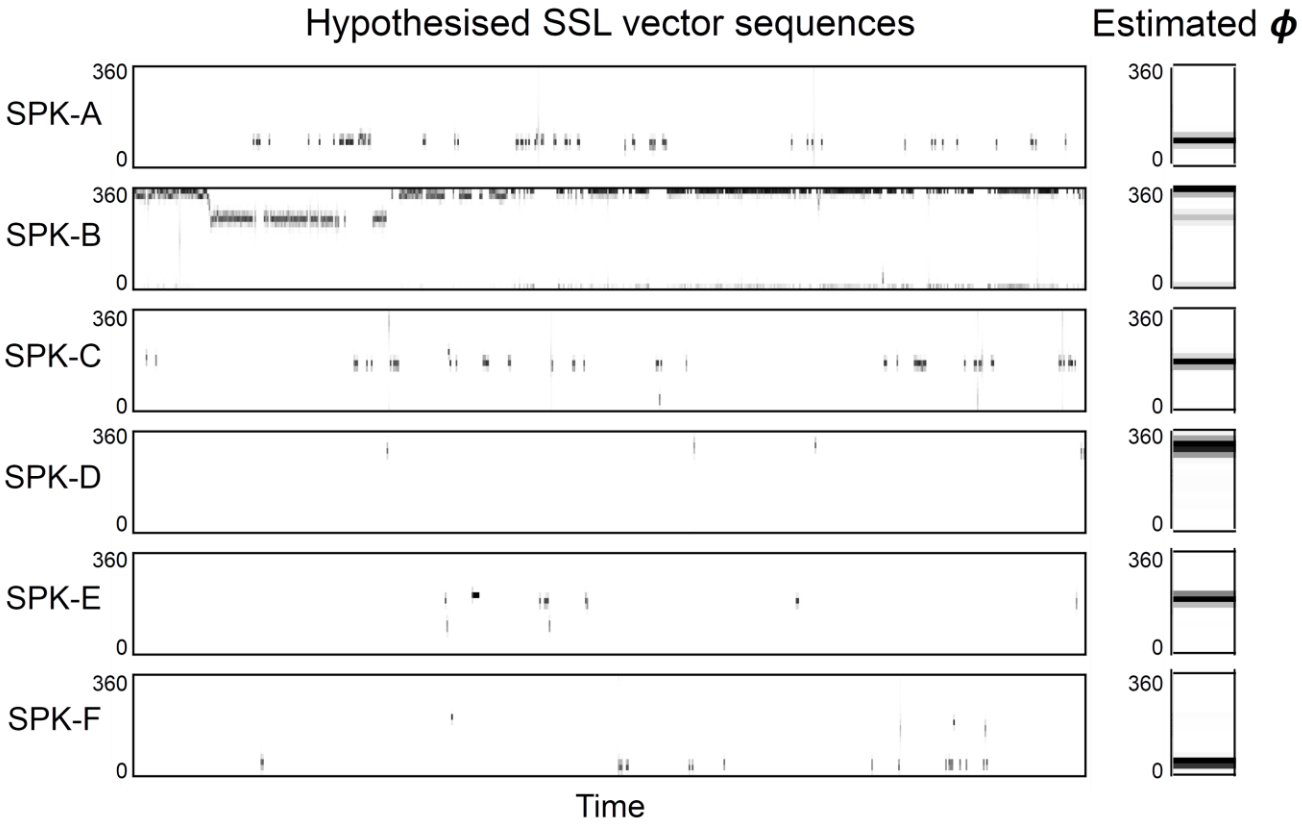
Microsoft

Experiments

Use SSL	E-M fine-tune	Speaker-attributed WER (%) *dev*	*eval*
no	none	22.75	21.41
no	λ, η	22.47	21.15
Uniformly initialize ϕ			
yes	λ, η, ϕ	21.62	20.42
Initalise ϕ from AHC hypothesis			
yes	λ, η	22.25	20.55
yes	λ, η, ϕ	21.61	20.37

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

Experiments



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

Microsoft

Summary

Proposed:

- Incorporate speaker location into HMM diarisation.

Results:

- Speaker location is complementary to speaker embeddings.

Future work:

- Investigate speaker movement tracking in diarisation.