

Unsupervised Fine-tuning of Speaker Diarisation Pipelines using Silhouette Coefficients

Lucas van Wyk (MuST, NWU)
Prof. Marelle Davel (MuST, NWU)
Dr. Charl van Heerden (SAIGEN)

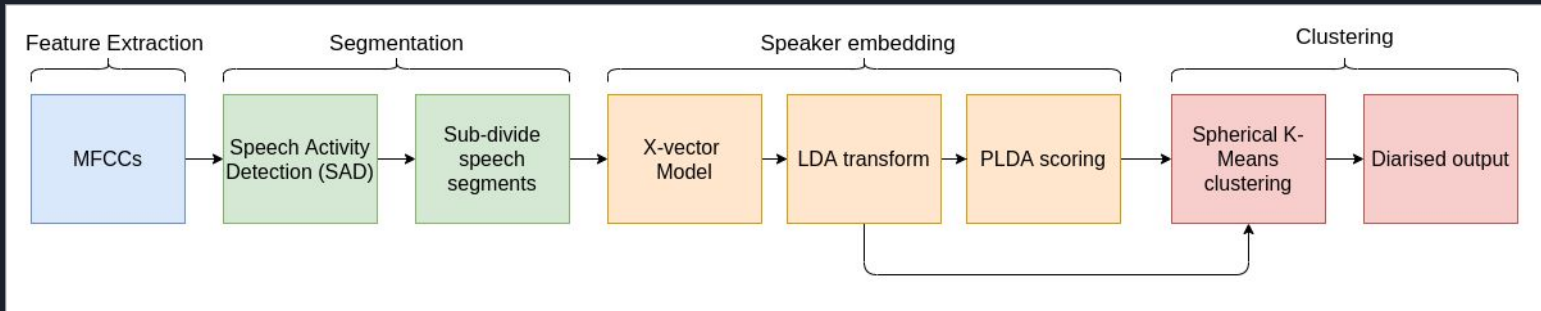


Introduction

1. What is speaker diarisation?
 - Who spoke when?
 - Timestamps and speaker label
2. Large amount of training data required
3. Domain-specific
4. How to adapt a pre-trained system to a new, low-resource domain and automate the domain adaptation process (unsupervised fine-tuning)?
5. ~20 hours of low-quality telephonic South African speech

Pre-trained baseline: Overview

- Out-of-domain data:
 - Voxceleb 1 (145,265 utterances/8,251 speakers)
 - Voxceleb 2 (1,092,009 utterances/36,237 speakers)
 - Data augmentation (RIRs & MUSAN)
- In-domain data:
 - 118 calls
 - 9 call centres
 - ~20 hours

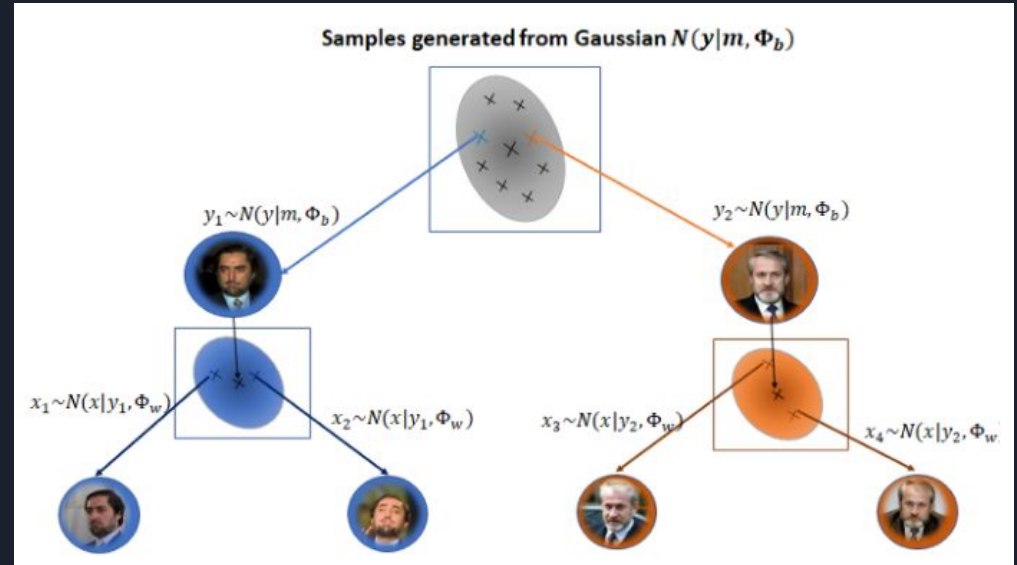


Domain-adaptation: PLDA

- The PLDA model:
 - Generative model (GMM)

$$P(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\mathbf{y}, \theta_W)$$

$$P(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{m}, \theta_B)$$





Domain-adaptation: Use of PLDA

- Inferring class identity
 - PLDA Score Matrix: Ratio of same/different speaker likelihoods

$$\begin{bmatrix} \mathbf{X}_{1,1}, \mathbf{X}_{2,1}, \mathbf{X}_{3,1}, \mathbf{X}_{4,1} \\ \mathbf{X}_{1,2}, \mathbf{X}_{2,2}, \mathbf{X}_{3,2}, \mathbf{X}_{4,2} \\ \mathbf{X}_{1,3}, \mathbf{X}_{2,3}, \mathbf{X}_{3,3}, \mathbf{X}_{4,3} \\ \mathbf{X}_{1,4}, \mathbf{X}_{2,4}, \mathbf{X}_{3,4}, \mathbf{X}_{4,4} \end{bmatrix}$$



Domain-adaptation: PLDA Interpolation

- Train PLDA and LDA on in-domain data
- Interpolate in-domain and out-of-domain PLDA
- Alpha is directly proportional to in-domain influence
- Optimal alpha determined using a small held-out set (supervised, per-corpus fine-tuning)
- We want to determine alpha automatically (unsupervised, per-file fine-tuning)

$$\theta_{W-adapt} = \alpha\theta_{W-in} + (1 - \alpha)\theta_{W-out} \quad \text{and}$$


$$\theta_{B-adapt} = \alpha\theta_{B-in} + (1 - \alpha)\theta_{B-out} \quad \text{with}$$

$$\alpha \in [0, 1]$$



The Silhouette Coefficient

- Measure of cluster quality based on
 - Intra-cluster distance (same-class)
 - Inter-cluster distance (different-class)
- Compute SC in two ways:
 - Use cosine distance between embedding (Standard SC)
 - Use columns of PLDA score matrix (Score Matrix SC)



Silhouette Coefficients for Domain Adaptation (unsupervised, per-file fine-tuning)

1. Train in-domain LDA and PLDA
2. Perform diarisation with PLDA interpolation on each file over a range of different values for alpha
3. Calculate SC for each alpha
4. For each file (call), identify the alpha with the highest SC



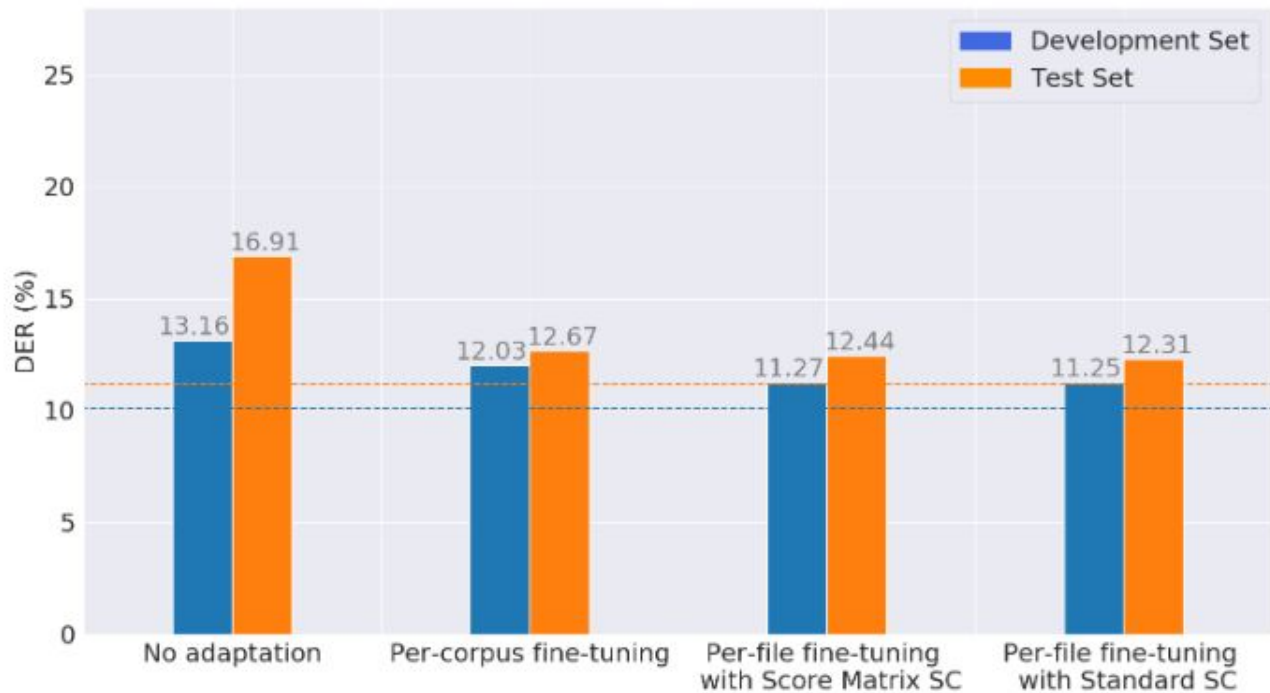
Experimental Setup: Training

- 80-20 Development-Test split of SATS corpus
- Train in-domain LDA and PLDA (development set)
- Perform interpolation for alpha in the range 0.5-1.0

	Calls	Hours
Call Centre 1	19	1.95
Call Centre 2	9	1.96
Call Centre 3	6	1.21
Call Centre 4	5	1.43
Call Centre 5	17	4.66
Call Centre 6	4	2.41
Call Centre 7	32	3.82
Call Centre 8	17	1.25
Call Centre 9	9	1.44
	118	20.13

- Estimate optimal alpha on development set (per-corpus, supervised)
- Estimate optimal alpha using SC (per-file, unsupervised)
- Compare per-corpus vs. per-file fine-tuning. (In terms of DER)

Experimental Results





Experimental Results

	Development Set		Test Set	
	DER	Relative	DER	Relative
Per-corpus fine-tuning, Oracle Speakers	12.03	8.59	12.67	25.07
Per-file fine-tuning, Score matrix SC	11.27	14.36	12.44	26.43
Per-file fine-tuning, Standard SC	11.25	14.51	12.31	27.20



Conclusion

- Possible to fine-tune the PLDA interpolation process in an unsupervised, per-file manner.
- Standard SC > Score Matrix SC (*for DER)
- SC ~ quality of adaptation



Questions?



Experimental Setup: Scoring

- Diarisation Error Rate (DER)
- 250ms Collar
- Oracle number of speakers

$$DER = \frac{\textit{False Alarm} + \textit{Missed Detection} + \textit{Speaker Confusion}}{\textit{Total Duration of Time}}$$

Experimental Results

