Segment-oriented evaluation of speaker diarisation performance

Rosanna Milner, Thomas Hain

Speech and Hearing Group, Department of Computer Science, University of Sheffield, UK

{rmmilner2,t.hain}@sheffield.ac.uk

Introduction

We propose a segment based F-measure, which specifically addresses issues such as reference errors, matching start and end boundaries, and speaker pairing. The performance of the metric is analysed in the context of state-of-the-art systems and compared with other existing metrics. It is shown to give a deeper insight into the segmentation quality over the standard metrics, and thus better value to understand impact on follow on tasks such as ASR.

Existing metrics for diarisation

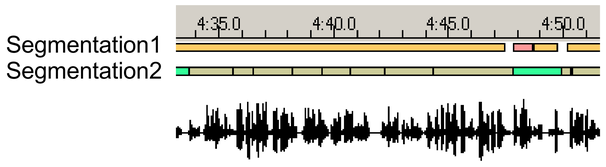
Diarisation error rate

Diarisation error rate (DER) considers missed speech, false alarm speech and speaker error:

$$DER = MS + FA + SE$$

Disadvantages:

- the collar around reference boundaries is typically 0.25 seconds, equivalent to 0.5 seconds around the boundary, represents at least a whole word and this time is not scored
- speaker mapping gives priority to large clusters and can ignore small clusters
- time based, does not consider segmentation quality



Boundary evaluation

Boundary methods using the F-measure and Dynamic Programming cost gives an average boundary error in time:

$$F = 2\frac{PRC * RCL}{PRC + RCL}$$

Disadvantages:

- deletions and insertions are treated equally
- DPC the metric will give most information if the units to be assessed are of approximately equal length
- does not consider what “type” of boundaries the matches are

Purity measures

Cluster purity describes the spread of speakers across a cluster and speaker purity describes the amount of clusters covered by a speaker. An overall purity calculation combines both cluster and speaker purity measures:

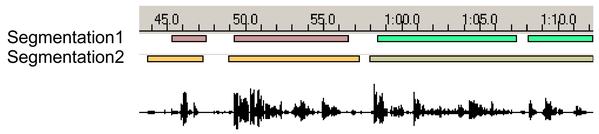
$$K = \sqrt{acp * asp}$$

Disadvantages:

- frame-based, does not evaluate segmentation

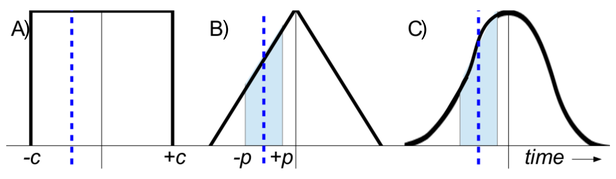
Segment F-measure

Matching start and end boundaries



Collar on reference boundary:

- allows for reference errors and uncertainty
- can be applied to reference boundary times (on either side) allowing for system boundaries to fall within this region
- equivalent to the assumption that the actual boundary is represented by a uniform probability density function (pdf) of certain width around the boundary
- estimate the probability of the hypothesis segment falling into a region using uniform (A), triangular (B) or Gaussian (C) distributions



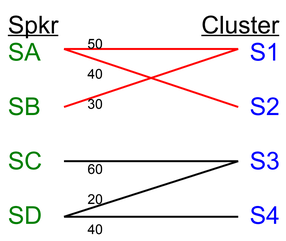
Padding on hypothesised boundary:

- allows for uncertainty in hypothesis
- larger probabilities introduces more leniency

Mapping speaker labels

Probability, or score, that a reference speaker, is mapped to hypothesised cluster, given all the observations:

- full search, all possible matchings and scores found
- for each speaker-cluster pair, cost is combination of all other pair scores
- combination of speaker-cluster pairs found which gives lowest cost



Highest pair score	
SA-S1	**50**
SA-S2	40
SB-S1	30

Lowest total cost	
SA-S1,SB-Ø	30+40=70
SA-S2,SB-S1	**50**

Multiple hypothesised segments

Non-overlapping:

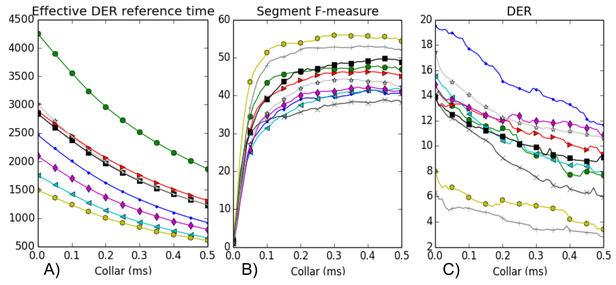
- smoothing where any adjacent segments with a limited gap are merged

Overlapping:

- the hypothesis segment with the matching boundaries and speaker label is chosen to be correct and the other hypothesised segments are considered as insertions

Evaluation

Using various collars, values from system BBC.4 individual files are shown for A) scored time used in DER, B) segment F-measure and C) DER.



File	u-sF	t-sF	g-sF	DER	DPC	bF	K
SAD							
RT07.1	1.5	1.9	2.2	3.7	2.9	25.4	-
RT07.2	1.2	1.3	1.4	7.2	15.8	19.8	-
BBC.3	0.1	0.2	0.2	9.9	1.4	79.2	-
BBC.4	72.4	73.6	74.8	2.0	0.2	94.5	-
DIA							
RT07.1	2.0	2.5	3.3	32.9	1.0	48.2	29.6
RT07.2	0.6	0.6	0.8	43.9	2.5	23.6	28.6
BBC.3	7.2	7.6	8.2	21.4	0.7	80.4	63.3
BBC.4	38.2	39.3	40.8	11.7	0.4	84.6	72.6
DIA - INDIVIDUAL FILES							
BBC.3x	5.2	5.9	6.1	17.9	0.9	14.2	66.5
BBC.4t	33.1	33.8	35.4	17.7	0.4	67.4	63.5

Results (collar 0.1 ms) show:

- metric is strict in finding matching segments
- different distributions allow for more lenient scoring
- other metrics can hide segmentation errors

Data	u-sF	t-sF	g-sF	DER	WER-MGB	WER-TED
RT07.1	2.0	2.5	3.3	32.9	31.5	27.1
RT07.2	0.6	0.6	0.8	43.9	30.2	28.1

Follow on automatic speech recognition (ASR) task for two systems:

- DER shows large gap whereas WERs are within 2%
- segment metric may be a clearer indication of the WERs

Conclusion

- DER (and other metrics) have shortcomings including lack of segmentation evaluation
- Proposed metric matches reference with hypothesised segments for deeper insight into speaker diarisation performance
- metric gives a more stable performance assessment and rank ordering of results

Download:
mini.dcs.shef.ac.uk/resources/sw/dia_segmentfmeasure

2016 IEEE International Conference on Acoustics, Speech, and Signal Processing, Shanghai, China, 20-25 March 2016