Segment-oriented evaluation of speaker diarisation performance

Rosanna Milner, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK
rmmliner2, t.hain@sheffield.ac.uk

Abstract

High performance diarisation is a necessity for a variety of applications, and the task has been studied extensively in the context of broadcast news and meeting processing. Upon introduction of the task in NIST led evaluations, diarisation error rate (DER) was introduced as the standard metric for evaluation, and it has been consistently used to compare systems ever since. DER is a frame based metric that does not penalise for producing many short segments. However, practical systems that require diarisation input are typically not able to cope well with such artefacts. In this paper we illustrate the need for an alternative metric focussing on segments, instead of duration or boundaries only. 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 for to understand impact on follow on tasks such as ASR.

Index Terms: speaker diarisation, diarisation error rate, boundary information, purity measures

1. Introduction

Speaker diarisation is an important task for audio indexing, and a prerequisite for other speech processing tasks such as automatic speech recognition (ASR) [1, 2]. The objective is to split the audio into speech segments which are associated with a single speaker, and to identify among the set of segments those that are spoken by the same speaker. The difficulty of the task is not only to group the speakers correctly, but also to find the correct number of clusters (i.e. speakers). Diarisation has been well studied over the years, research has been performed on telephone [3], meeting [4] and broadcast media data [5], for example. Several toolkits are available in the public domain for this task, however most are designed to perform well for a specific type of data [6, 7, 8].

NIST [4] established the task and the diarisation error rate (DER) [9] for use in the speaker diarisation evaluations, conducted during the years 2002-9. It has been widely adopted to be the standard metric for evaluating systems, and is based on detecting missed speech, false alarms and speaker error in terms of time only. Alternative methods for assessment of diarisation also exist, such as boundary centric methods that focus on evaluating the segmentation stage, such as the Dynamic Programming Cost [10]. The F-measure has also been used to evaluate the number of inserted, deleted and matched boundaries [11]. The clustering stage can be evaluated using speaker and cluster purity measures [12]. The DER metric obscures several significant properties of system outputs that are relevant for practical tasks, and therefore the search for alternative metrics is an important research question.

One issue arises from vagueness of what constitutes a segment. Typically, references are created by humans who will choose pauses where it is semantically meaningful. Therefore sentences (or “spurts” [13]) can be seen more as semantic units. This is done for a reason – it makes no sense to listen to fragmented sentences for a person. Similarly, downstream applications such as translation or summarisation require semantically meaningful fragments. DER avoids that issue by using frame level correctness rather than segmental correctness, allowing for completely fragmented output without any penalty.

The second weakness of DER is that it does not allow for ambiguity in reference and output. As even for manual labellers it is not completely clear where boundaries have to be placed decisions need to be lenient and allow for correctness ranges, for example in the form of confidence on boundary location. DER does not accomodate this and therefore leniency is expressed by deletion of data, as further outlined below. Hence highly conversational speech becomes easier to detect although in fact the segments are harder to find and overlap plays a big role.

Both weaknesses have to do with the lack of decision orientation in assessing diarisation output. For this reason we propose a metric based on the F-measure, a measure of accuracy, in terms of segments [1].

2. Existing metrics for diarisation

The DER is the standard and most commonly used evaluation metric. Others include the DPC and boundary F-measure which evaluates segment boundaries, and speaker and cluster purity measures which evaluate the speaker clustering.

2.1. Diarisation error rate

DER measures the amount of time not accurately assigned to speech, a specific speaker or non-speech, and is widely adopted across the field [1, 2, 9]. It is calculated using the equation:

\[
DER = \frac{\sum_{s=1}^{S} dur(s)(max(N_{\text{ref}}(s), N_{\text{hyp}}(s)) - N_{\text{correct}}(s))}{\sum_{s=1}^{S} dur(s)N_{\text{ref}}}
\]

where \(S\) is the number of speaker segments, in which the reference and the system output file contain the same speaker pair, and \(dur(s)\) is the length of a segment. The \(N_{\text{ref}}\) and \(N_{\text{hyp}}\) represent the numbers of speakers in the reference and hypothesis segment and \(N_{\text{correct}}\) is the amount of correctly matched speakers [14]. It is simply the sum of missed time error (MS), false alarm error (FA) and speaker error (SE). Missed speech refers to reference speech detected as silence, false alarm is reference silence detected as speech, and speaker error measures the percentage of scored time in which a speaker label is assigned to the wrong speaker. A “collar” around the reference boundaries

\[\text{http://mini.dcs.shef.ac.uk/resources/sw/dia_segmentfmeasure} \]
excludes that region from scoring, thus showing the uncertainty in the reference annotation.

There are several disadvantages to the DER. Firstly, the use of the collar is problematic. The standard of 0.25 seconds is equivalent to 0.5 seconds around the boundary. Assuming 3 words a second, this is at least one whole word. Furthermore, data is removed from scoring. As will be illustrated later on, this can amount to half of the overall data. Secondly, the reference speakers are mapped to hypothesised speaker labels by selecting the mapped pair with the maximum amount of coinciding time. This gives priority to large clusters and can ignore small clusters. The third and arguably biggest issue is that the number of segments do not feature in the metric. This implies that either the introduction of short inter-segment gaps or the bridging of short gaps hardly get penalised. In Figure 1, multiple segments have been hypothesised for one reference segment, and if reference speaker SA is mapped to hypothesised speaker S1) there is a segment with an incorrect speaker label. However, as the majority of the reference speech has been found and has the correct speaker mapped label, the DER will be a reasonable result. It measures frame by frame instead of error based on correctly detected speech segments [15].

2.2. Boundary evaluation
Segment boundaries are important information on segmentation. The Dynamic Programming Cost (DPC), as defined in [10], calculates sequences of boundary information (the reference and the hypothesis output) using the absolute time difference between the two as a cost. DPC is measured in milliseconds per reference boundary, and is found by dividing the cost by the number of reference boundaries. An F-measure can be calculated which gives a score involving the number of matched, inserted and deleted boundaries in terms of precision (PRC) and recall (RCL) [11]. Precision refers to when a true boundary is matched and recall refers to when a hypothesis boundary correctly corresponds to a boundary in the reference:

$$\text{PRC} = \frac{N_{\text{mat}}}{N_{\text{mat}} + N_{\text{ins}}}, \quad \text{RCL} = \frac{N_{\text{mat}}}{N_{\text{mat}} + N_{\text{del}}}$$

$$F = 2 \frac{\text{PRC} \times \text{RCL}}{\text{PRC} + \text{RCL}}$$

where $N_{\text{mat}}$, $N_{\text{ins}}$, and $N_{\text{del}}$ are the number of matches, insertions and deletions respectively.

A problem with this boundary evaluation is that deletions and insertions are treated equally. Arguably in a speaker diarisation system it is worse to produce misses than false alarms, as these are unrecoverable portions of speech. As for the DPC, the metric will give most information if the units to be assessed are of approximately equal length. However, for diarisation this is often not the case.

This method does penalise split segments in terms of increasing the number of insertions, but it does not consider what “type” of boundaries the matches are. For example, looking to the left and the right of the boundary, it could be NONSPEECH-SPEECH, SPEECH-NONSPEECH or SPEECH-SPEECH (different speakers, referred to as a speaker change). It finds the closest boundary in time without checking the type of boundary.

Figure 2 shows an example of a match for the second reference boundary but it should be considered incorrect due to the types. However, the metric is easily changed to penalise any matches which do not have the same type of boundary and this updated boundary F-measure is used for the rest of this paper.

2.3. Purity measures
Purity measures are usually used for general clustering algorithms but can be applied to speaker clustering in the form of cluster purity and speaker purity. Cluster purity describes how a cluster is contained to only one speaker and speaker purity describes how well a speaker is constricted to only one cluster. They do not give detailed information of the segmentation performance. They are described in more detail in [12] where $n_i$ is the number of frames in cluster $i$, $n_j$ is the number of frames uttered by speaker $j$, $n_{ij}$ is the frame count in cluster $i$ spoken by speaker $j$, $N$ is the cluster count, $N_s$ is the number of speakers and $N_f$ is the number of frames. Cluster purity, $p_c$, of cluster $i$ and the average cluster purity, $acp$, are:

$$p_c = \sum_{i=k}^{N} \frac{n_i^2}{n_i^2 + n_i^2}, \quad acp = \frac{1}{N} \sum_{i=1}^{N} p_c n_i$$

Secondly, the speaker purity, $p_j$, of speaker $j$ and average speaker purity, $asp$, are:

$$p_j = \sum_{i=1}^{N_s} \frac{n_{ij}^2}{n_j^2}, \quad asp = \frac{1}{N_s} \sum_{j=1}^{N_s} p_j n_j$$

An overall purity calculation combines both cluster and speaker purity measures:

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

which is used as a method to evaluate different systems.

Speaker and cluster purity is again frame based and describes the spread of speakers across clusters and vice versa. It does not show the user whether the audio has been segmented correctly, meaning it must be used alongside another metric to evaluate the segmentation.

3. Segment F-measure
As outlined above, existing metrics only allow to focus on very specific aspects while ignoring others. In this work we propose to use complete segment match as the base, where a segment is correct if it matches the boundaries of the reference. Similar to the methods for boundary detection (outlined in §2.2), precision and recall, and consequently F-measures can be used to assess performance. Such a segment-oriented metric allows to address the issues raised with other metrics.

Performance is evaluated in terms of matched segments and each reference segment is treated individually. A hypothesised segment is matched to a reference segment if its start and end boundaries lie within reach of the reference segment’s start and end boundaries, and the speaker labels are equivalent. It compensates for small errors in references, and there are three different approaches to allow for boundary leniency. The metric also includes a segment-based speaker mapping method and deals with overlapping segments.
matches, a segment-based speaker mapping is chosen. A time-
based method is not used as a very long but incorrectly clustered
segment can lead to suboptimal assignment.

Instead of the greedy search, a full search is implemented
in order to find the globally optimal mapping of reference
to hypothesised speakers. In a first step, all possible matchings
between reference speakers and hypothesised clusters and their
scores are found. Next, for every pair with a score found, the
possible combinations of other pairs of speakers are found and
the scores of any ignored pairs are counted as a cost, or er-
ror, for this combination of pairs. Finally, the combination of
speaker and label pairs which produce the lowest cost is cho-
sen to be the correct speaker mappings. Figure 4 illustrates how
this improves over methods based on greedy search. The greedy
method would select SA-S1 to be correct as it has the highest
score, removing these two labels from further mappings mean-
ning both SB and S2 would be unmatched labels (and thus both
are an error). However, full search looks at all combinations and
costs: where SA-S1 pairing would have a cost of 30 + 40 = 70
with two unmatched speakers, and the alternative would be SA-
S2 with cost 50 and SB-S1 with cost 0, an overall cost of 50 +
0 = 50 making it the more optimal combination.

3.4. Multiple hypothesised segments

It can happen that multiple hypothesised segments can be asso-
ciated with the same reference segment. If they are not overlap-
ing then smoothing is carried out. Any adjacent segments with
a limited gap can be merged (smoothing as on the reference).
If this produces a single segment with matching boundaries and
equivalent speaker label then it is considered a match.

If 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. If
more than one segment matches with boundaries, the segment
with the equivalent speaker label is chosen to be correct.

4. Evaluation

In this section we compare results for the segment F-measure
(sF) with DER, DPC, boundary F-measure (bF) and K, the over-
all purity measure. We evaluate across two data domains each
using two speaker diarisation systems. Speaker diarisation can
be a prerequisite for tasks such as ASR, so a good understanding
of the segmentation quality is vital.

4.1. Data and systems

The first dataset, RT07, is single channel meeting data. It con-
sists of 35 speakers across 8 meetings recorded in four different
meeting rooms and was collected for the NIST Rich Transcrip-
tion 2007 evaluation [16]. It has been used with two different
Deep Neural Network (DNN) based systems (RT07.1, RT07.2).
DNN segmentation followed by adaptation using a pre-trained
DNN to separate speakers [17, 18]. The second is a media
broadcast programme from the BBC where there is always four
speakers, a host and three guests. It has been used with SHoUT
[8] which uses an unsupervised model training regime (BBC.3)
and a system using DNN segmentation and alignment on the in-
dividual head microphone (IHM) channels, resulting in an sin-
Results

The sF metric can be used to evaluate speech activity detection (SAD) as well as speaker diarisation (DIA). For SAD, before any preprocessing of merging adjacent segments with equivalent speakers, the speaker labels are removed (treated as a single one, “speech”) and any two segments which overlap are treated as one. This will of course give higher scores in comparison to the DIA scores as the speaker labels are ignored.

Overall scores for SAD and DIA are shown in Table 1. The overall sF scores are found by weighting each file by the number of reference segments. There is a clear difference between the sF and the other metrics. For SAD, three systems achieve similar sF scores whereas the DER is not correlated. The same is true for DIA, the two RT07 systems achieve similar sF scores but the DER varies by 10%. This backs our argument that the DER is misleading in terms of segmentation evaluation. The BBC.3 system provides poor segmentation shown in the sF, 0.4%, however the DPC and bF both fail here giving improved scores, 0.7 ms and 80.4% respectively. The purity, K, also fails to show the poor segmentation.

The uniform (u-sF), triangular (t-sF) and Gaussian (g-sF) distributions at boundaries are used for adding leniency when matching. The thresholds for t-sF and g-sF were tuned separately and the padding applied is 0.01 seconds, 20 ms around the hypothesised boundary. For both SAD and DIA, the t-sF greatly improves the scores allowing for large leniency, for example, for DIA RT07.1 increases from 55.5% to 75.6% and the poor performing BBC.3 system increases from 0.4% to 5.8%. The g-sF improves for DIA in a much smaller degree and in some SAD cases, it drops slightly, e.g. BBC.4 goes from 76.5% for the u-sF to 74.7%. In the final section of Table 1, scores for individual files with roughly the same sF of 17% are shown. The sF scores differ in these cases, highlighting differences not observable in other metrics, particularly the DER, as their evaluation does not consider the segmentation quality.

Figure 5 displays the effects on u-sF and DER scores for RT07.1 and BBC.4 under a changing collar. The middle plot is the amount of scored time evaluated in the DER, as the collar increases more time is being removed and not evaluated on. For most files this means a reduction of data by more than 50% at the standard collar of 0.25 seconds. Arguably difficult, highly variable sections of the data have been removed at this point. As a consequence of the data change the DER itself drops, by up to 20% absolute, and in some cases halves the error.

Table 1: Overall SAD and DIA scores for the four systems, including selected individual files, where u-sF, t-sF, and g-sF are the sF scores using a uniform, triangular and Gaussian boundary distribution respectively. All scores use a 0.1 second collar.

<table>
<thead>
<tr>
<th>File</th>
<th>u-sF</th>
<th>t-sF</th>
<th>g-sF</th>
<th>DER</th>
<th>DPC</th>
<th>bF</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT07.1c</td>
<td>39.5</td>
<td>71.0</td>
<td>30.2</td>
<td>17.8</td>
<td>0.3</td>
<td>69.8</td>
<td>57.0</td>
</tr>
<tr>
<td>RT07.2d</td>
<td>58.5</td>
<td>82.9</td>
<td>60.9</td>
<td>17.2</td>
<td>0.2</td>
<td>76.6</td>
<td>75.9</td>
</tr>
<tr>
<td>BBC.3x</td>
<td>0.4</td>
<td>7.1</td>
<td>0.7</td>
<td>17.9</td>
<td>0.9</td>
<td>14.2</td>
<td>66.5</td>
</tr>
<tr>
<td>BBC.4t</td>
<td>33.1</td>
<td>41.9</td>
<td>35.3</td>
<td>17.7</td>
<td>0.4</td>
<td>67.4</td>
<td>63.5</td>
</tr>
<tr>
<td>DIA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT07.1c</td>
<td>55.3</td>
<td>75.6</td>
<td>63.4</td>
<td>10.5</td>
<td>0.3</td>
<td>86.6</td>
<td>68.6</td>
</tr>
<tr>
<td>RT07.2d</td>
<td>55.7</td>
<td>79.6</td>
<td>57.5</td>
<td>21.9</td>
<td>0.2</td>
<td>84.2</td>
<td>73.9</td>
</tr>
<tr>
<td>BBC.3x</td>
<td>0.4</td>
<td>5.8</td>
<td>0.7</td>
<td>21.4</td>
<td>0.7</td>
<td>80.4</td>
<td>63.3</td>
</tr>
<tr>
<td>BBC.4t</td>
<td>38.6</td>
<td>46.6</td>
<td>40.6</td>
<td>11.7</td>
<td>0.4</td>
<td>84.6</td>
<td>72.6</td>
</tr>
</tbody>
</table>

The authors would like to thank Jana Eggink and the BBC. This work was supported by the EPSRC Programme Grant EP/I031022/1 Natural Speech Technology (NST).

Figure 5: Using various collars, values from RT07.1 individual files are shown for A) segment F-measure, B) DER and C) scored time used in DER. Plots D) and E) show the uniform segment F-measure and DER scores respectively for the BBC.4 system.
6. References


