Segment-oriented evaluation of speaker diarisation performance

Rosanna Milner, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK

rmmilner2, 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 obfuscates 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:

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

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_{ref} and N_{hyp} represent the numbers of speakers in the reference and hypothesis segment and N_{corr} 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

[1] http://mini.dcs.shef.ac.uk/resources/sw/dia_segmentfmeasure

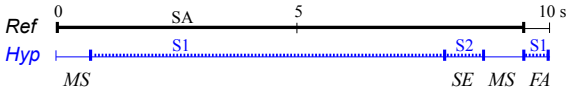

Figure 1: *Example of DER scoring, where MS, FA and SE are missed speech, false alarm and speaker error time segments. There is no hypothesised segment which represents the reference yet there is only small MS, FA and SE and thus low DER.*

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 gets 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], aligns 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:

$$PRC = \frac{N_{mat}}{N_{mat} + N_{ins}}, \quad RCL = \frac{N_{mat}}{N_{mat} + N_{del}} \tag{2}$$

$$F = 2\frac{PRC * RCL}{PRC + RCL} \tag{3}$$

where N_{mat}, N_{ins} and N_{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.

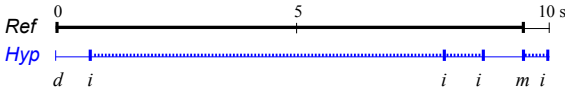

Figure 2: *Example of DPC and F-measure scoring. This shows the initial hypothesised boundary as a deletion, d, several inserted boundaries, i and an incorrectly matched reference end boundary, m.*

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_c is the cluster count, N_s is the number of speakers and N is the number of frames. Cluster purity, $p_{i.}$, of cluster i and the average cluster purity, acp, are:

$$p_{i.} = \sum_{j=1}^{N_s} \frac{n_{ij}^2}{n_{i.}^2}, \quad acp = \frac{1}{N}\sum_{i=1}^{N_c} p_{i.} n_{i.} \tag{4}$$

Secondly, the speaker purity, $p_{.j}$, of speaker j and average speaker purity, asp, are:

$$p_{.j} = \sum_{i=1}^{N_c} \frac{n_{ij}^2}{n_{.j}^2}, \quad asp = \frac{1}{N}\sum_{j=1}^{N_s} p_{.j} n_{.j} \tag{5}$$

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

$$K = \sqrt{acp * asp} \tag{6}$$

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.

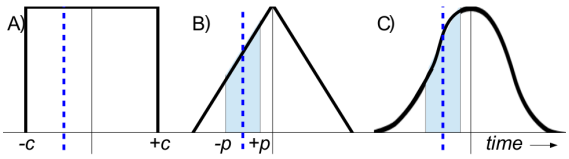

Figure 3: *Three different distributions used for boundary matching: A) uniform, B) triangular and C) Gaussian. The vertical solid line represents the reference boundary and the vertical dashed line represents where the hypothesised boundary has fallen, and c and p are the collar and padding respectively.*

3.1. Reference errors

Reference timings can be manually created or come from alignment of a transcript. Either method would produce a certain amount of discrepancy between the timings found and the true timings. In DER, this is controlled by applying a collar around the reference boundaries where the score is ignored, leading to loss of scoring time. The segment based method allows to define a range in which the boundary should fall. The range value (also collar) is thus an expression of reference uncertainty without loss in scoring power. Initially, adjacent reference segments with a limited time gap and same speaker labels can be merged as a smoothing method.

3.2. Matching start and end boundaries

For each reference segment, a hypothesis segment is to be found with equivalent start and end times. As mentioned, a collar can be applied to the reference boundary times (on either side) allowing for the hypothesis boundaries to fall within this region. This is equivalent to the assumption that the actual boundary is represented by a uniform probability density function (pdf) of certain width around the boundary. Consequently, one can estimate the probability of the hypothesis segment falling into a region using uniform or other distributions. The probabilities for start and end boundaries can be multiplied and a threshold used to decide whether the segment matches or not.

Distributions tested include uniform, triangular and Gaussian and are depicted in Figure 3. The collar represents the width, or variance for the Gaussian distribution. An optional padding variable can be applied which allows for larger probabilities and introduces more leniency. For the uniform case padding would only turn the decision into a different effective boundary and hence is reduced to a simple match or not.

3.3. Mapping speaker labels

NIST scoring pairs reference and hypothesised speakers based on time, by mapping the speaker and label with the overall maximum time matched until all reference speakers have an equivalent hypothesised label if possible. The proposed metric only considers reference speakers and hypothesised labels which occur in segments with matching start and end boundaries. Reference segments without matches are ignored for this stage as the speaker labelling may not be reliable. For example, if reference speakers S_A and S_B exist and hypothesised speaker labels S_1 and S_2 exist, the probability, or score, that a reference speaker, S_A, is mapped to hypothesised label, S_1, given all the observations can be expanded:

$$P(r = S_A, h = S_1|O) = P(h = S_1|r = S_A, O)P(r = S_A|O) \quad (7)$$

Both parts could be represented in different ways, either by time (amount of coinciding time between matched segments) or the number of segments. As this F-measure is based on segment 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.

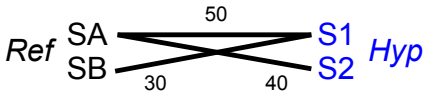

Figure 4: *Example of sub optimality of greedy speaker mapping. The optimal solution (SA-S2,SB-S1) gives a lower cost.*

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 error, for this combination of pairs. Finally, the combination of speaker and label pairs which produce the lowest cost is chosen 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 meaning 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 associated with the same reference segment. If they are not overlapping 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 overall 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 consists of 35 speakers across 8 meetings recorded in four different meeting rooms and was collected for the NIST Rich Transcription 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 individual head microphone (IHM) channels, resulting in an sin-

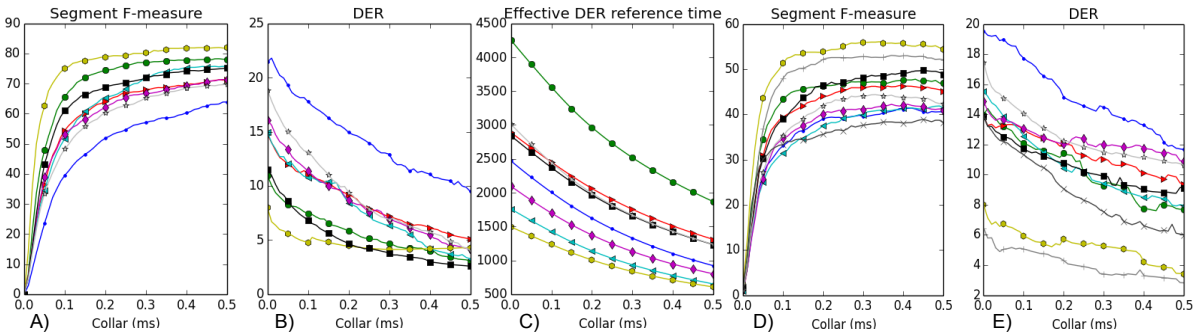

Figure 5: *Using various collars, values from RT07.1 individual files are shown for A) segment F-measure, B) DER and C) scored time used in DER. Plots D) and E) show the uniform segment F-measure and DER scores respectively for the BBC.4 system.*

gle distant microphone (SDM) output (BBC.4).

4.2. 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 76.3% for the u-sF to 74.7%. In the final section of Table 1, scores for individual files with roughly the same DER 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. One can further observe that for DER the rank ordering between different

File	u-sF	t-sF	g-sF	DER	DPC	bF	K
SAD							
RT07.1	76.6	85.2	80.8	2.6	0.2	94.8	-
RT07.2	74.7	79.8	74.3	14.7	0.3	90.6	-
BBC.3	0.4	5.0	0.2	9.9	1.4	79.2	-
BBC.4	76.3	78.9	74.7	2.0	0.2	94.5	-
DIA							
RT07.1	55.5	75.6	63.4	10.5	0.3	86.6	68.6
RT07.2	55.5	79.6	57.5	21.9	0.2	84.2	73.9
BBC.3	0.4	5.8	0.7	21.4	0.7	80.4	63.3
BBC.4	38.6	46.6	40.6	11.7	0.4	84.6	72.6
DIA - INDIVIDUAL FILES							
RT07.1c	39.5	71.0	50.2	17.8	0.3	69.8	57.6
RT07.2d	58.5	82.9	60.9	17.2	0.2	76.6	75.9
BBC.3x	0.4	7.1	0.7	17.9	0.9	14.2	66.5
BBC.4t	33.1	41.9	35.3	17.7	0.4	67.4	63.5

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.*

data elements can change, in the case of BBC.4 quite considerably. In contrast, for sF both for RT07.1 and BBC.4 the rank ordering remains stable throughout collar change. Furthermore, the error values become stable much sooner, at approximately the precision level that the reference was generated. Rank ordering is important for comparison of systems as the smoothing parameters should not affect the outcome of comparisons.

5. Conclusion

This paper presented an overview of diarisation assessment metrics and provided an analysis of their strength and weaknesses. The shortcoming of the key metric, diarisation error rate, was demonstrated in experimental results on different tasks and with different systems. Amongst others the main shortcoming is insensitivity to segmentation errors. We have introduced a new metric that incorporates both assessment of speaker error and highlights issues with segmentation. The result is a more stable performance assessment and rank ordering of results, which will allow for more meaningful assessment of diarisation performance in conjunction with other downstream tasks.

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).

6. References

[1] S. E. Tranter, "Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio," *ICASSP*, vol. 1, pp. 1013–1016, 2006.

[2] X. M. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Trans. Audio Speech and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[3] "NIST: Speaker Recognition Evaluation," http://www.nist.gov/speech/tests/spk/, accessed: 11-01-2016.

[4] "NIST: Rich Transcription Evaluation Project," http://www.itl.nist.gov/iad/mig/tests/rt/, accessed: 11-01-2016.

[5] T. Hain and P. C. Woodland, "Segmentation and classification of broadcast news audio," in *ICSLP'98*, 1998, pp. 851–854.

[6] D. Vijayasenan and F. Valente, "DiarTk: An Open Source Toolkit for Research in Multistream Speaker Diarization and its Application to Meetings Recordings." *INTERSPEECH*, pp. 5–8, 2012.

[7] M. Rouvier, P. Gay, E. Khoury, T. Merlin, S. Meignier, and L. Mans, "A Free State-of-the-art Toolbox for Broadcast News Diarization," *INTERSPEECH*, 2013.

[8] M. A. H. Huijbregts, "Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled," Ph.D. dissertation, 2008.

[9] "Diarisation error rate scoring code, NIST," http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl, accessed: 11-01-2016.

[10] V. Z. van Vuuren, L. ten Bosch, and T. Niesler, "A Dynamic Programming Framework for Neural Network-based Automatic Speech Segmentation," *INTERSPEECH*, pp. 2287–2291, 2013.

[11] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Process. Lett.*, vol. 11, no. 8, pp. 649–651, 2004.

[12] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM." *Proc. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 573–576, 2002.

[13] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation," in *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7, 2001*, 2001, pp. 1359–1362.

[14] X. A. Miro, "Robust Speaker Diarization for meetings," Ph.D. dissertation, Universitat Politecnica de Catalunya, 2006.

[15] J. M. Pardo, "Speaker Diarization Features: The UPM Contribution to the RT09 Evaluation," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 2, pp. 426–435, 2012.

[16] "NIST: Rich Transcription Evaluation 2007," http://nist.gov/speech/tests/rt/2007/index.html, accessed: 11-01-2016.

[17] R. Milner, O. Saz, S. Deena, M. Doulaty, R. W. M. Ng, and T. Hain, "The 2015 sheffield system for longitudinal diarisation of broadcast media," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, 2015.

[18] O. Saz, M. Doulaty, S. Deena, R. Milner, R. Ng, M. Hasan, Y. Liu, and T. Hain, "The 2015 sheffield system for transcription of multi–genre broadcast media," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, 2015.