Studying Human-Based Speaker Diarization and Comparing to State-of-the-Art Systems

Simon W. McKnight, Aidan O. T. Hogg, Vincent W. Neo, Patrick A. Naylor
Department of Electrical and Electronic Engineering
Imperial College London, UK
{s.mcknight18, a.hogg, vincent.neo09, p.naylor}@imperial.ac.uk

***Abstract*—Human-based speaker diarization experiments were carried out on a five-minute extract of a typical AMI corpus meeting to see how much variance there is in human reviews based on hearing only and to compare with state-of-the-art diarization systems on the same extract. There are three distinct experiments: (a) one with no prior information; (b) one with the ground truth speech activity detection (GT-SAD); and (c) one with the blank ground truth labels (GT-labels). The results show that most human reviews tend to be quite similar, albeit with some outliers, but the choice of GT-labels can make a dramatic difference to scored performance. Using the GT-SAD provides a big advantage and improves human review scores substantially, though small differences in the GT-SAD used can have a dramatic effect on results. The use of forgiveness collars is shown to be unhelpful. The results show that state-of-the-art systems can outperform the best human reviews when no prior information is provided. However, the best human reviews still outperform state-of-the-art systems when starting from the GT-SAD.**

I. Introduction

A. Speaker Diarization Background

Speaker diarization is the process of distinguishing different speakers in any given speech signal and identifying the times during which they speak. It involves two fundamental aspects: (i) segmentation of speech data into either constant time periods (e.g. a fixed number of frames) or non-constant time periods that are homogeneous in some way (e.g. single speaker speech, overlapping speaker speech or no speech); and (ii) clustering and/or labelling the segments identified to attribute them to individual speakers [1], [2], [3], [4].

Good speaker diarization has many important applications, such as being used as a first step before applying automatic speech recognition (ASR), thereby making existing audio transcripts more meaningful and searchable, or to assist hearing impaired people with identifying different speakers on conference calls. It is the focus of much academic research and several challenges (e.g. DIHARD I, II and III [5], [6], [7], CHiME-6 [8] and VoxSRC 2021 Track 4 [9]). This paper focuses on single channel recordings.

B. Labels and Scoring

By far the most commonly used speaker diarization scoring mechanism is the diarization error rate (DER) using the `md-eval.pl` file originated for the NIST Rich Transcription challenges held from 2002 to 2009 [10]. This is a time-based measure that involves comparing the system outputs to a ground truth reference file. As shown in this paper, results can be dramatically affected by small differences in the ground truth, which means it is not a wholly satisfactory scoring mechanism. However, accurate scoring is an essential element of determining the best systems and what the best approaches for improvement are.

A distinction needs to be drawn between speaker diarization systems that start from the ground truth speech activity detection (GT-SAD) from those that do not (they either (a) use a separate speech activity detection (SAD) system either as a pre-processing step [11], [12] or as a post-processing step [13] or (b) have it built directly into their system in some way such as end-to-end systems [14]). The DIHARD challenges operate separate scoreboards for these – Track 1 covers systems that start from the GT-SAD, Track 2 covers systems that do not, and error rates for the latter are considerably higher. For example, for the DIHARD III core set, the Track 1 winning DER was 13.45% and the Track 2 winning DER was 19.37%, a difference of 5.92% [15]. Moreover, this difference increases steadily going down the list – for example, the average of top five Track 1 DERs was 14.766% and five Track 2 DERs was 21.836%, a difference of 7.07%. Similarly, in DIHARD II the Track 1 winning DER was 18.42% and the Track 2 winning DER was 27.11%, a difference of 8.69%.

Scoring uses the ground truth labels (GT-labels). When a GT-SAD is used, it is created from GT-labels, so is only as accurate as those GT-labels and has the same assumptions.

A final reason why accurate labelling is necessary is that if inaccurate labels are used in supervised model training, they could result in the models themselves being confused. For example, training a supervised model using labels that show someone speaking when in fact they are not or *vice versa* will harm the model, particularly if it is not robust to those uncertainties/errors. Systems such as [11] mitigate this by training features on data that do not have inaccurate labels (specifically x-vectors on VoxCeleb 1 and 2 data [16]) and extract features from comparatively long speech sections of 1.5 s, before refining systems with probabilistic linear discriminant analysis (PLDA) models trained on potentially less accurate labels in the validation set [12], [17].

C. Human Reviews Experiment Structure

There are three distinct parts:

1) ***Experiment 1*** - no prior information given to reviewers, so they need to decide start and end times of each label

as well as distinguishing the speakers. This is consistent with how Track 2 of the DIHARD challenges work;

2) ***Experiment 2*** - reviewers start from the GT-SAD, so they need to label one or more speakers at various times within the GT-SAD only. This is consistent with how Track 1 of the DIHARD challenges work; and
3) ***Experiment 3*** - reviewers start from the blank GT-labels, so they just need to distinguish speakers. This is not done in any current challenges, but is helpful for identifying human ability to discriminate speakers purely based on what they can hear rather than more subjective placement of timing boundaries.

Reviews were done in this order for individual reviewers so more detailed prior information from earlier reviews would not influence later reviews.

II. Human Reviews Analysis

The time-based scoring methodology used by `md-eval.pl` [10] and this paper is

$$DER = \frac{\tau_M + \tau_{FA} + \tau_{SE}}{\tau_{TOTAL}}, \quad (1)$$

where τ_M is the miss time, τ_{FA} is the false alarm time, τ_{SE} is the speaker error time, τ_{TOTAL} is the total speech time and DER is the diarization error rate (expressed as a percentage). The individual components of DER are $MISS$, FA and SE, which are the relevant times τ_M, τ_{FA} and τ_{SE} respectively divided by τ_{TOTAL} and expressed as a percentage.

Scoring of diarization systems is highly sensitive to assumptions made and small variations. For example, if a speaker is speaking but has a short pause between two parts (regardless of whether they are in the same sentence), then one ground truth label might treat as a single utterance whereas another splits it into two. Diarization challenges explicitly state the minimum pause duration before an utterance is split into two, such as 300 ms for the NIST Rich Transcription challenges [10] or 200 ms for the DIHARD challenges [7], but these are hard to maintain consistently and there is considerable subjectivity. Even more problematically, most labelling readily available is much less accurate, so ideally a way could be found to take advantage of the less accurate labels without penalising the scoring of more accurately labelled systems.

Uniform forgiveness collars are crude attempts to mitigate this problem [18]. These exclude collars of plus and minus the collar size around the GT-labels from scoring. Simple illustrative examples were set up for 10 s sections with 9 s speech (with variations for pauses and overlapping speech) and are shown in Figs 1 and 2. As shown in Fig. 1, the forgiveness collars around the GT-label starts and ends will ignore system predictions in that collar (it "forgives" errors, but equally does not reward correct predictions in those collars), but if imprecise ground truth labelling is used then system labels showing pauses where the ground truth does not will be penalised. Conversely, Fig. 2 shows that precise ground truth labels along with short overlapping speech results in excessive forgiveness and relatively little scored speech. Some diarization challenges still use collars (e.g. VoxSRC 2021 Track 4 [9]), but most do not (e.g. DIHARD I, II and III [5], [6], [7] and CHiME-6 Track 2 [8]) and some researchers have opined that collars should be excluded [12].

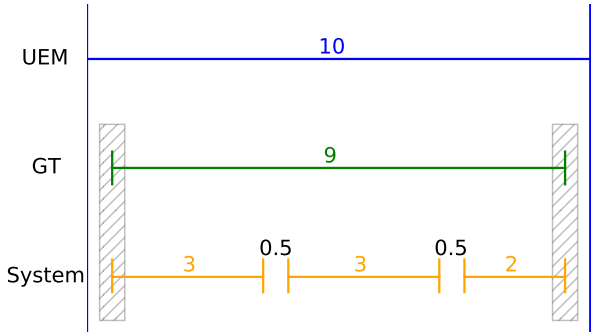

Fig. 1: Example segmentation and collars with imprecise ground truth labelling and precise system labelling (250 ms collar: $DER = 11.8\%$; no collar: $DER = 11.1\%$). All numbers in s. UEM means unpartioned evaluation map.

III. Human Reviews Experimental Design and Results

A. Datasets and Systems Used

A five-minute extract of the AMI Corpus [19] ES2008a headset recordings was used, specifically between 0:30.000 and 5:30.500 (slightly longer than five-minutes to complete the last utterance). This is recorded in the unpartitioned evaluation map (UEM) file in scoring; note that the UEM is simply the portion of the speech file to be evaluated [10].

The GT-SAD used in Experiment 2 is created from the GT-labels. Initially, the GT-labels were constructed from the `ES2008a.[A-D].segments.xml` files (GT1), but it was found that (a) these contained silence of 0.25-0.5 s at the starts and ends of each segment which greatly increased miss errors [20] (possibly on the basis that false positives were seen as worse than false negatives in a SAD used for ASR [21], [22]) and (b) they included non-lexical sounds such as laughter and speech. Consequently, for Experiment 1 the scoring was subsequently repeated by excluding GT-labels (and consequently

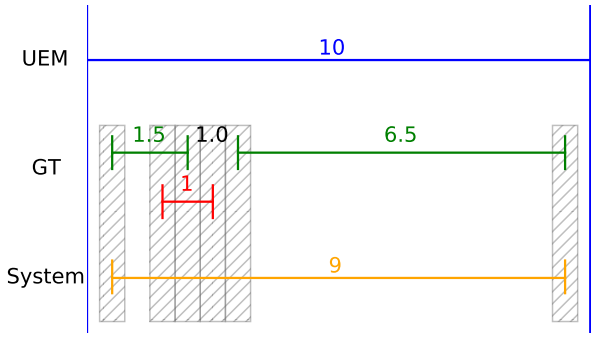

Fig. 2: Example segmentation and collars with precise ground truth labelling, overlapping speech and imprecise system labelling (250 ms collar: $DER = 0\%$; no collar: $DER = 16.7\%$). All numbers in s.

TABLE I: Comparison of ES2008 extracts statistics for different ground truths. "Tot." is the total aggregate speech time (i.e. adding together all overlapping speech), "Dur." means duration, "Num." means number, "Ave." means average, "Seg." means segment, "Comb." refers to the combined speech time (i.e. the time during which there is at least one speaker) and "Ch. Rate" is the rate at which the speakers change.

Segments	Tot. Dur. (s)	Tot. Num. Segs	Comb. Dur. (s)	Comb. Num. Segs	Overlap (%)	Ch. Rate (Hz)	Ave. Seg. Dur. (s)
GT1	251.05	47	231.35	30	8.52	0.406	5.34
GT2	247.00	40	230.66	29	6.15	0.327	6.12
GT3	230.12	41	220.32	31	4.45	0.356	5.61
GT4	237.05	48	221.59	32	6.98	0.319	6.98

the GT-SAD constructed from it) that only contained non-lexical sounds (GT2), and for Experiment 2 the GT-SAD was reconstructed from GT2. After that, GT-labels and GT-SAD were constructed from the `ES2008a.[A-D].words.xml` files generated for the AMI corpus for words only using forced alignment and HTK [19] (conveniently already extracted in the "only_words" directory of [12], [23]) (GT3) and lastly constructed from those same AMI corpus files but this time including non-word vocal sounds and conveniently in the "word_and_vocalsounds" directory of [12], [23] (GT4 and, together with GT1, GT2 and GT3, the GTs). References to GT-labels and GT-SAD generated from specific ground truths are GT1-labels and GT1-SAD, for example.

Some results for Experiment 1 are shown for all four GT-labels to highlight sensitivity to specific labels, but Experiments 2 and 3 are based on GT3 only. Table I highlights differences across these different GT-labels and Fig. 3 shows a histogram of the GT3-labels durations.

For Experiment 2, the starting point GT-SAD and the scoring GT-labels were both based on GT3. For Experiment 3, the blank starting point GT-labels and the scoring GT-labels were both based on GT3.

The reviewers listened to the speech file individually using Sennheiser HDA 300 headphones following the instructions at [24]. The labels were recorded on Audacity [25] with the experiment administrator present to ensure the experiments were carried out consistently and to avoid inadvertent errors (e.g. if the end of a label was selected that moved the start as well, or if pressing `ctrl-z` for "undo" caused unexpected changes). Audacity was clearly not the perfect tool for labelling and all reviewers found it awkward to use.

The reviewer numbers for Experiment 1 are different to those in Experiments 2 and 3 because the numbering was based on the order that the reviewers carried out the Experiments. Experiment 1 was carried out several weeks before Experiments 2 and 3 (these two were on the same day).

The anonymous results of the experiments are available at [24]. A Colab notebook analysing them is at [26]. Sections III-B to III-D highlight the most significant results.

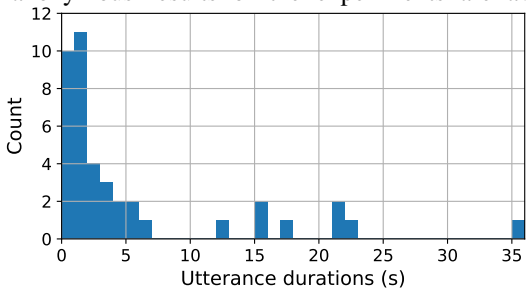

Fig. 3: Histogram of GT3-labels durations.

B. Experiment 1 Results

With 13 human reviews, Fig. 4 shows the timeline of the GT3 for the speech extract along with each of the reviews. The error scored against GT3 for each reviewer plotted against the number of speech segments predicted is shown in Fig. 5. There are two outliers, one of whom predicted few segments and the other predicted many (and the latter with a very low total speech time), so these are ignored in Table II.

Fig. 5 shows the importance of predicting roughly the same number of segments as the ground truth labels, regardless of whether forgiveness collars are applied. Including the two outliers and ignoring the system results, this graph looks strongly parabolic with minimum around the number of ground truth labels. That said, there is a clear vertical line where a number of reviewers predicted much the same number of segments (8 of the 13 reviewers predicted 35 to 38 segments) yet their DERs differed by as much as 5.86%.

Table II shows that excluding non-lexical sounds such as laughter and coughing improves results (GT2 better than GT1 and GT3 better than GT4), which is not surprising as the reviewers had been told to exclude them. Using GT3 rather than GT2 or GT4 rather than GT1 improves DER results significantly, highlighting the fact that small differences can make a dramatic difference to the results, regardless of what forgiveness collars are used. Using collars increases DER standard deviations slightly in all cases, which is counter-intuitive as similar reviews before applying collars should become almost identical if those collars are properly applied, which is not the case here. Uniform forgiveness collars only work well if the predicted number of segments matches the number used by the relevant ground truth and they are consistent in where the speech and non-speech portions are.

Comparing to state-of-the-art systems was rather cumbersome. Many of them require a good SAD, but using some of the industry standard pre-trained models such as Google VAD [27], [28] or Silero [29] that could not be tuned on

TABLE II: Experiment 1 means and standard deviations (STDs) for different GTs with either (a) 250 ms collars or (b) no collars. All figures in %.

GT	250 ms Means	250 ms STDs	0 ms Means	0 ms STDs
GT1	11.93	1.51	18.94	1.43
GT2	11.02	1.46	17.20	1.45
GT3	8.95	1.60	15.60	1.53
GT4	10.27	1.66	17.62	1.44

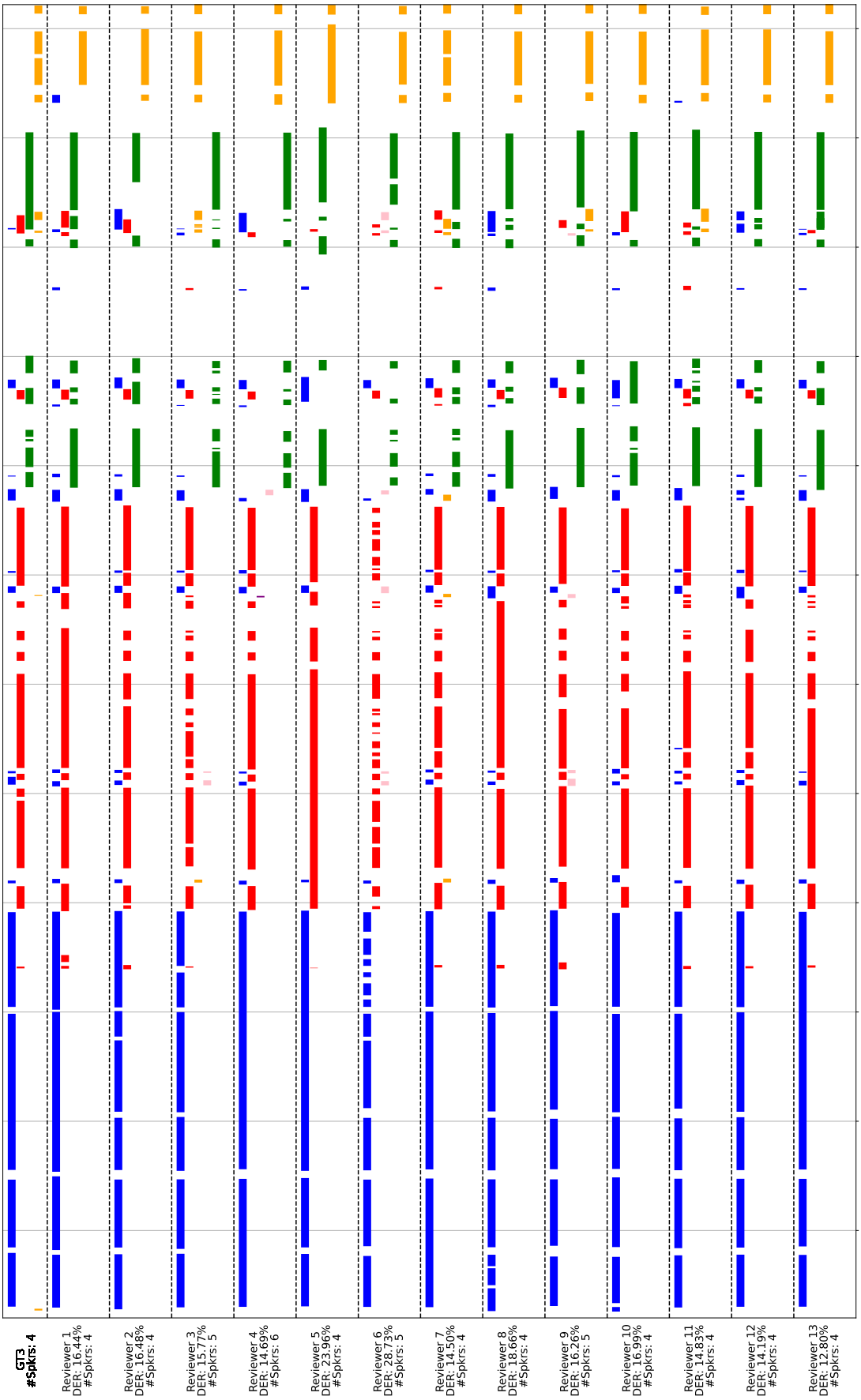

Fig. 4: Experiment 1 predictions against GT3. The GT3 colours are FEE029 in blue, FEE030 in red, MEE031 in green and FEE032 in orange. The reviewer predictions are matched to those based on the greatest length of time matched by `md-eval.pl`, with additional unmatched predicted speakers in pink and purple. No forgiveness collar was applied to calculate the DER. "Spkrs" refers to the total number of speakers predicted by that reviewer.

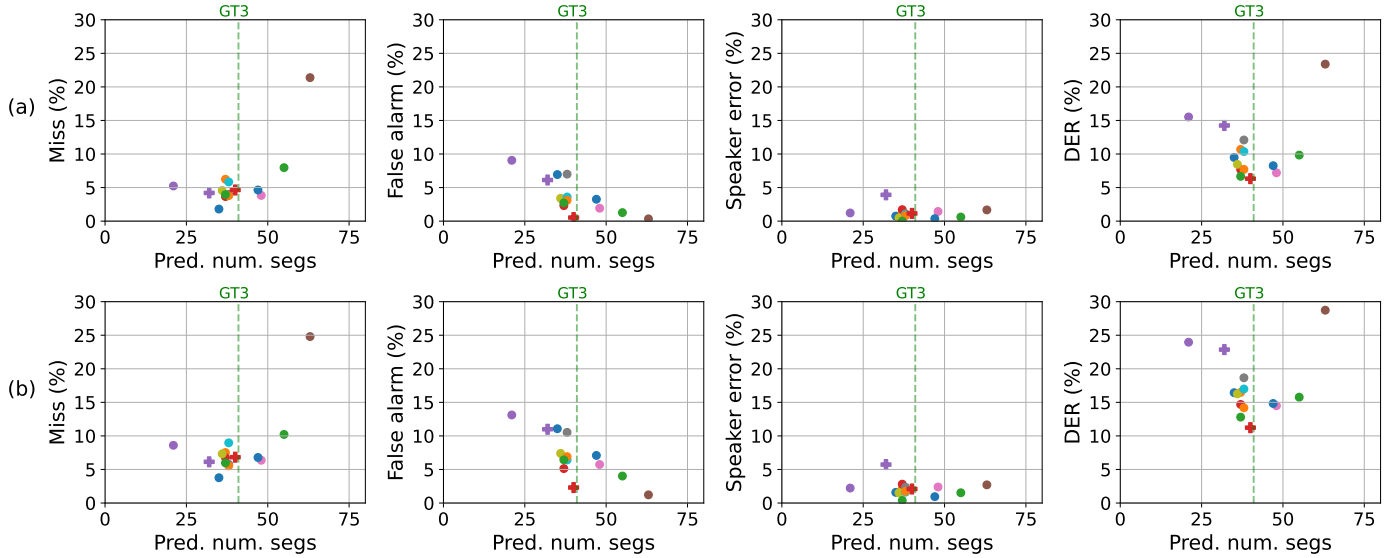

Fig. 5: Experiment 1 graphs showing reviewer errors in dots plotted against the number of segments they predicted for (a) 250 ms forgiveness collars and (b) no forgiveness collars. The `pyannote.audio` V2 baseline is in + and V1 in +.

the validation set led to much worse results than the human reviews. These systems are considered in Experiment 2 as the GT-SAD improves their results significantly. Others such as end-to-end systems either required some tuning on a validation set [14] and/or did not have an option to start from the GT-SAD. The baseline system tested in Experiment 1 is the default pre-trained model for `pyannote.audio` Version 2.0.1 (V2) [30], [31], which had a no-collar DER of 11.23% that outperformed even the best human review of 12.80%. It predicted 40 speaker segments, close to the 41 of GT3, so is consistent with the observation that segment number predictions closer to that of the GT used improves performance. The `dia_ami` pre-trained model for `pyannote.audio` Version 1.1.2 (V1) [32] was also tested as this model had the option of starting from the GT-SAD and is used in Experiment 2. The `dia_dihard` pre-trained model for `pyannote.audio` V1 was also tried and predicted 99 segments with 28.15% DER (not shown in any of the graphs), highlighting sensitivity of model performance to training on suitable data.

Fig. 6 illustrates the trade-off between misses and false alarms. Increasing the miss generally leads to a similar decrease in the false alarm, and *vice versa*. Ignoring the two outliers, a 45° diagonally descending line would be a decent fit. Fig. 7 shows the importance of predicting roughly the right aggregate length of time as the ground truth labels. This is not as clearly defined as the parabola in Fig. 5 is though, suggesting that predicting the right number of segments is more important than predicting the right overall speech time.

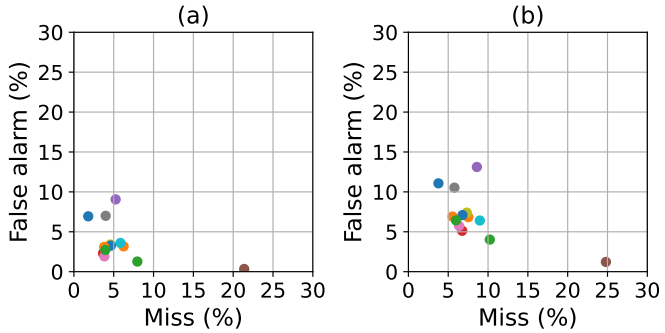

Fig. 6: Experiment 1 graphs of miss v. false alarm trade-off for (a) 250 ms collars and (b) no collars.

C. Experiment 2 Results

With 10 human reviews, the DER means and standard deviations shown in Table III are significantly better than for Experiment 1 (11.11% better without a collar, 6.92% better with). No outliers for Experiment 2 need to be excluded. This confirms that it is far easier to diarize speakers if the GT-SAD (or SAD that accurately reflects the GT-SAD) is used. This is consistent with results from speaker diarization challenges; for example, (a) in DIHARD II the system SAD Track 2 winning DER was 27.11% and the GT-SAD Track 1 winning DER was 18.42% [33] and (b) in DIHARD III the system SAD Track 2 winning DER was 19.27% and the GT-SAD Track 1 winning DER was 13.45% [15]. If a single speaker was predicted at all times, FA would fall to zero and $MISS$ would reflect the missed overlapping speakers. Fig. 8 shows there is much less importance in predicting roughly the same number of segments as the ground truth labels than for Experiment 1, which is because the GT-SAD takes away the biggest source of uncertainties and consequently reduces the variance.

Longer utterances contain more speaker information, and reviewers found them easy. Only 2 of 10 reviewers correctly identified the short 0.41 s utterance of "hmm" at the start as being by a different speaker from the immediately following speaker. In Experiment 1, only 3 of 13 reviewers counted that

TABLE III: Experiment 2 means and STDs for either (a) 250 ms collars or (b) no collars. All figures in %.

GT	250 ms Means	250 ms STDs	0 ms Means	0 ms STDs
GT3	2.03	0.64	4.49	0.73

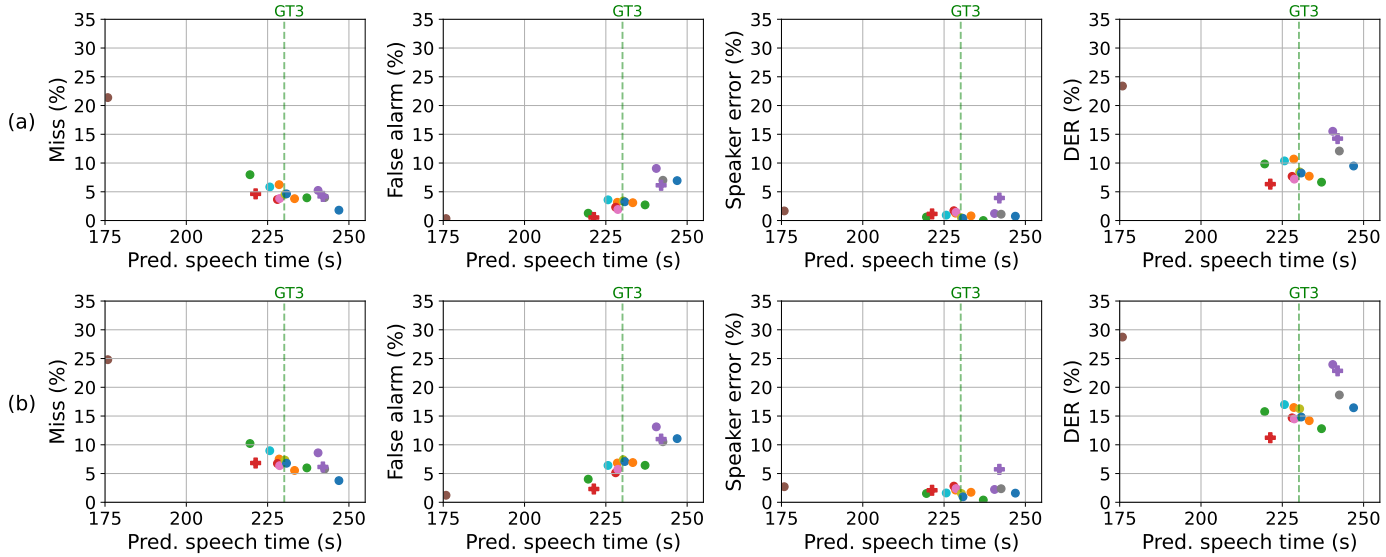

Fig. 7: Experiment 1 graphs showing the reviewer errors in dots plotted against the aggregate total speech time they predicted for (a) 250 ms forgiveness collars and (b) no forgiveness collars. The `pyannote.audio` V2 baseline is in + and V1 in +.

short utterance speech and none identified it as being by a different speaker than the immediately following speaker.

The three state-of-the-art comparison systems are [11] (BDII), [12] (ResNet101) and the `pyannote.audio` V1 `dia_ami` pre-trained model. The `pyannote.audio` V1 and V2 systems use features generated from raw audio using SincNet [34] and have the significant advantage of being trained on AMI data. BDII and ResNet101 use x-vectors trained on VoxCeleb along with PLDA models trained on VoxCeleb and DIHARD validation set data.

D. Experiment 3 Results

Fig. 9 uses review times on the x-axis to distinguish the reviewers as the predicted number of segments and aggregate predicted speech time was identical for all. With 10 human reviews, these scores are dramatically better than those for Experiments 1 and 2. $MISS$ and FA naturally fall to zero, so the only errors come from SE which fall to 1.41% without collars and 0.68% with 250 ms collars as shown in Table IV.

E. Reviewer Observations

While the recordings were generally clear, the reviewers found the heavy breathing annoying. This was caused by the headset microphones being directly in front of the speakers' mouths rather than to the side, and may have affected the ability to hear quiet or whispered speech at times.

Several reviewers noted the female speakers all had similar pitch, so reviewers used semantic information to distinguish them at times rather than vocal pitch or timbre. Two reviewers who were non-native English speakers felt they were at a disadvantage compared to the native English speakers. Furthermore, times when an existing female speaker interjected in a higher-pitched voice or showing more emotion were often incorrectly thought to have been a different speaker altogether.

TABLE IV: Experiment 3 means and STDs for either (a) 250 ms collars or (b) no collars. All figures in %.

GT	250 ms Means	250 ms STDs	0 ms Means	0 ms STDs
GT3	0.68	0.69	1.41	1.03

All reviewers coped well when there were two speakers at the same time, but all struggled when there were more than two. Part of the problem was that overlapping speech tended to be very short, which often did not contain enough speaker information for the reviewer to determine who it was. Furthermore, reviewers were told to classify vocal sounds such as "hmm", "em", "um" and "uh" as speech (these are vocal disfluencies or filler words that generally do not contain semantic information, but do have some speaker information and could convey approval or disapproval of a point) but not non-lexical sounds such as laughter or coughing, and in some cases it was not easy to distinguish them. GT3 records 22.69 s of those utterances, a significant 9.04% of overall speech time.

IV. Discussion and Conclusion

Sections III-B to III-D make it clear that human reviews can produce wildly diverging performance, showing how sensitive the scoring methodology is to specific assumptions and variations. The main difficulties are caused by short utterances, especially any uncertainty as to whether they constitute proper vocal sounds that should be recorded, along with overlapping speech and discretion on whether short pauses in utterances should be treated as a break in those utterances.

The human reviews results show that the use of uniform forgiveness collars is somewhat arbitrary and generally not helpful. They make sense if the speech segments identified match those of the ground truth except for small deviations at the ground truth segment boundaries, but fail to take account of differences elsewhere. Applying forgiveness collars to the human reviews resulted in a significant decrease in every DER, as expected, but also led to an increase in the DER standard deviations. Each reviewer clearly has a different amount of error within the forgiveness collar, and reviewers with more

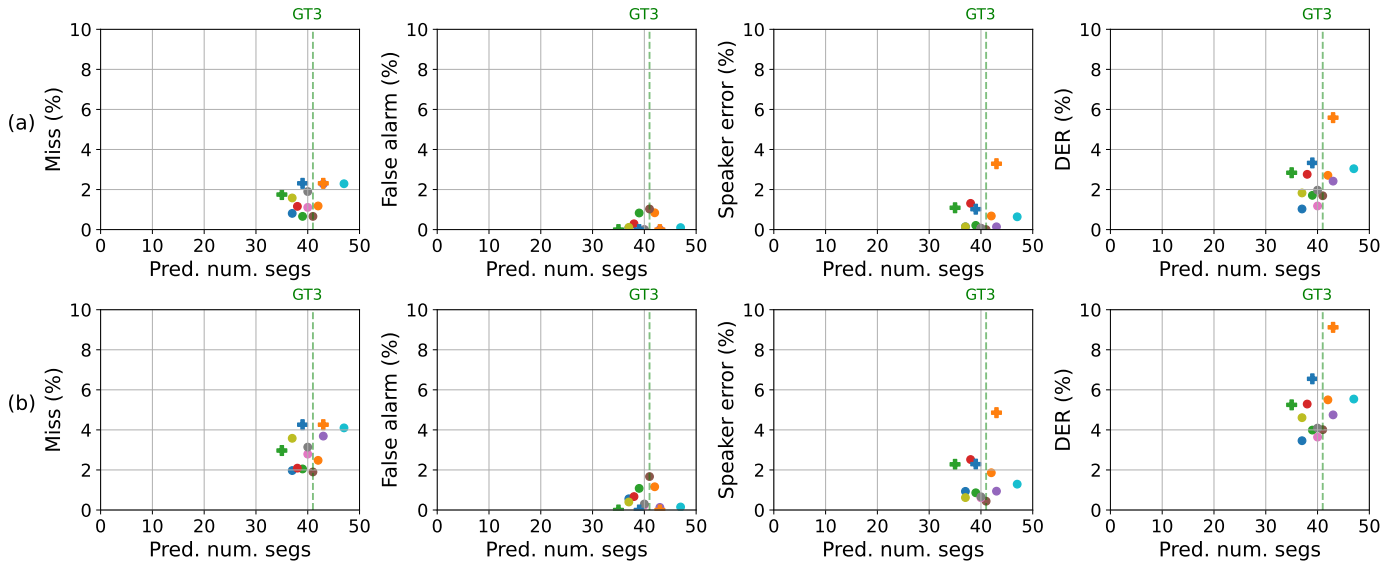

Fig. 8: Experiment 2 graphs showing reviewer errors in dots plotted against number of segments predicted using GT3-SAD for (a) 250 ms forgiveness collars and (b) no forgiveness collars. BDII in +, ResNet101 in +, `pyannote.audio` V1 in +.

errors in the forgiveness collars will have their DERs reduced by more than those with less when the forgiveness collars are excluded. If the overall errors were primarily due to imprecise placement of the speaker segments starts and ends within the forgiveness collars, the expectation is that reviewers with higher DERs would have more errors in the forgiveness collars than those with lower DERs, and consequently those higher DERs would be reduced by more than the lower DERs, resulting in a lower standard deviation in the DERs. As the standard deviation in fact increased, it means that reviewers with higher DERs did not in general have more errors in the forgiveness collars, and consequently must have had more errors elsewhere. However, had the ground truth made different assumptions about the pauses, those other reviews might have fared better when the forgiveness collars were excluded. All the forgiveness collar has done is make the DERs look better on the whole, it has not improved the results of all reviews equally or fairly as the ones that made different assumptions about the pauses are more harshly penalised.

In Experiment 1, `pyannote.audio` V2 outperformed all human reviewers, with no-collar DER 11.23% compared to 12.80% of the best human. In Experiment 2, 7 of 10 human reviews outperformed all baseline systems, with the best human no-collar DER of 3.46% compared to 5.25% of the best system. Also, `pyannote.audio` V1 dramatically improved from 22.86% DER in Experiment 1 to 5.25% in Experiment 2, highlighting sensitivity to SAD reflecting the ground truth. These results suggest that state-of-the-art diarization systems are already better than humans at predicting speaker segment times that are consistent with the ground truth (there are some

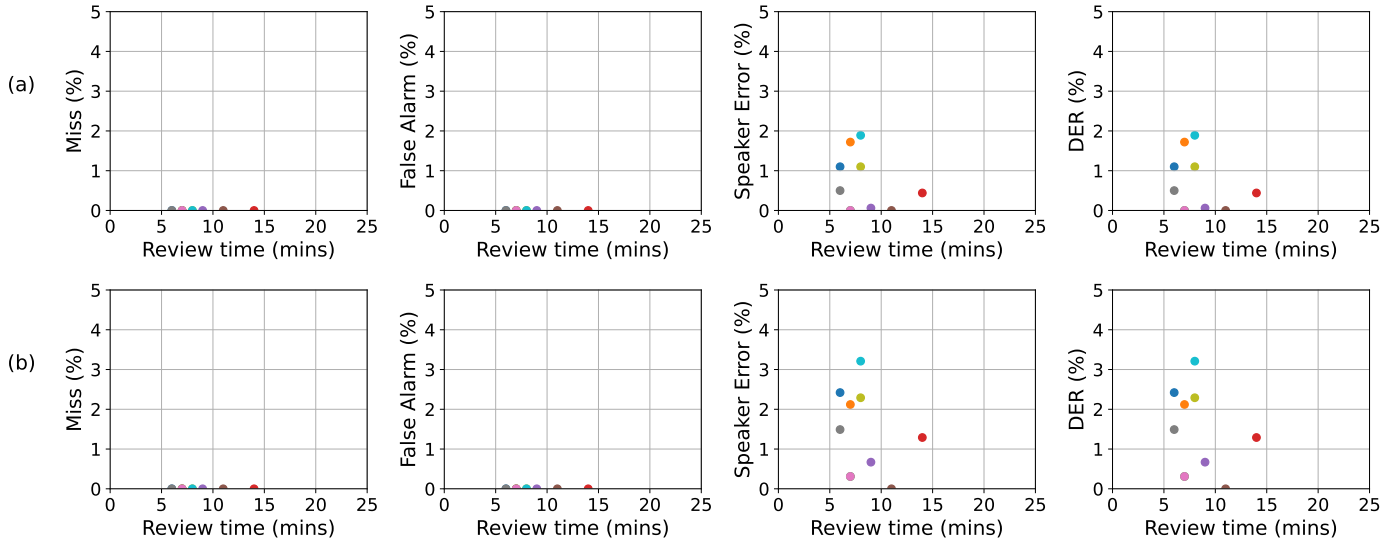

Fig. 9: Experiment 3 graphs showing reviewer errors in dots plotted against their review time based on the GT3-labels for (a) 250 ms forgiveness collars and (b) no forgiveness collars.

caveats here – it may have been fortunate to have chosen an Experiment 1 baseline system that had similar assumptions to the ground truth). However, good human reviews still outperform relatively simple state-of-the-art diarization systems when starting from GT-SAD, so humans are better at distinguishing speakers. More sophisticated systems involving multiple components, along with a SAD consistent with that used for scoring, are essential for system performance to reach and exceed human performance.

REFERENCES

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: recent advances with deep learning," *Comput. Speech and Language*, vol. 72/101317, 2022.

[2] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: a review of recent research," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 356–370, 2012.

[3] X. Anguera Miro, "Robust speaker diarization for meetings," Ph.D. dissertation, Universitat Politecnica de Catalunya, 2006.

[4] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 5, pp. 1557–1565, 2006.

[5] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," Tech. Rep., 2018. [Online]. Available: https://zenodo.org/record/1199638#.XkABaWj7Q2w

[6] ——, "The second DIHARD diarization challenge: dataset, task, and baselines - version 1.2," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 978–982.

[7] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third DIHARD challenge evaluation plan," 2020. [Online]. Available: https://dihardchallenge.github.io/dihard3/docs/third_dihard_eval_plan_v1.2.pdf

[8] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: tackling multispeaker speech recognition for unsegmented recordings," *arXiv:2004.09249 [cs, eess]*, 2020.

[9] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, "VoxSRC 2021: the third VoxCeleb speaker recognition challenge," 2022. [Online]. Available: https://arxiv.org/pdf/2201.04583.pdf

[10] NIST, "The 2009 (RT-09) rich transcription meeting recognition evaluation plan," 2009. [Online]. Available: https://web.archive.org/web/20100606041157if_/http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf

[11] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotný, H. Zeinali, and J. Rohdin, "BUT system for the second DIHARD speech diarization challenge," 2020. [Online]. Available: http://arxiv.org/abs/2002.11356

[12] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks," *Comput. Speech and Language*, vol. 71/101254, 2020.

[13] Y. Wang, M. He, S. Niu, L. Sun, T. Gao, X. Fang, J. Pan, J. Du, and C.-H. Lee, "USTC-NELSLIP system description for DIHARD-III challenge," *arXiv:2103.10661 [cs, eess]*, 2021. [Online]. Available: http://arxiv.org/abs/2103.10661

[14] S. Horiguchi, N. Yalta, P. Garcia, Y. Takashima, Y. Xue, D. Raj, Z. Huang, Y. Fujita, S. Watanabe, and S. Khudanpur, "The Hitachi-JHU DIHARD III system: competitive end-to-end neural diarization and x-vector clustering systems combined by DOVER-lap," *arXiv:2102.01363 [cs, eess]*, 2021. [Online]. Available: http://arxiv.org/abs/2102.01363

[15] N. Ryant, "The third DIHARD speech diarization challenge - results," 2021. [Online]. Available: https://dihardchallenge.github.io/dihard3/results.html

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 2616–2620.

[17] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 3954, pp. 531–542.

[18] S. W. McKnight, A. O. T. Hogg, and P. A. Naylor, "Analysis of phonetic dependence of segmentation errors in speaker diarization," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2020.

[19] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: a pre-announcement," in *Proc. of the 2nd Int. Workshop on Mach. Learning for Multimodal Interaction (MLMI'05)*, 2006, pp. 28–39.

[20] J. Moore, M. Kronenthal, and S. Ashby, "Guidelines for AMI speech transcriptions," Tech. Rep., 2005. [Online]. Available: https://groups.inf.ed.ac.uk/ami/corpus/Guidelines/speech-transcription-manual.v1.2.pdf

[21] V. W. Neo, S. Weiss, S. W. McKnight, A. O. T. Hogg, and P. A. Naylor, "Polynomial eigenvalue decomposition-based target speaker voice activity detecction in the presence of competing talkers," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

[22] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[23] F. Landini, J. Profant, M. Diez, and L. Burget, "AMI diarization setup," 2020. [Online]. Available: https://github.com/BUTSpeechFIT/AMI-diarization-setup

[24] S. W. McKnight, "GitHub site with human review results," 2022. [Online]. Available: https://github.com/swm1718/HumanReviews

[25] "Audacity." [Online]. Available: https://www.audacityteam.org/

[26] S. W. McKnight, "Colab site analysing human review results," 2022. [Online]. Available: https://tinyurl.com/4ys4ba7t

[27] J. Wiseman, "Python interface to the WebRTC voice activity detector," 2016. [Online]. Available: https://github.com/wiseman/py-webrtcvad

[28] Google, "WebRTC: real-time communication for the open web platform," 2011. [Online]. Available: https://webrtc.org/

[29] A. Veysov, "Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier," 2021. [Online]. Available: https://github.com/snakers4/silero-vad

[30] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 7124–7128.

[31] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2021, pp. 3111–3115.

[32] H. Bredin, "Neural speaker diarization with pyannote.audio," 2020. [Online]. Available: https://github.com/pyannote/pyannote-audio

[33] N. Ryant, "The second DIHARD speech diarization challenge - results," 2019. [Online]. Available: https://dihardchallenge.github.io/dihard2/#results

[34] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," Aug. 2019. [Online]. Available: http://arxiv.org/abs/1808.00158