Analysis of Phonetic Dependence of Segmentation Errors in Speaker Diarization

Simon W. McKnight, Aidan O. T. Hogg and Patrick A. Naylor
Department of Electrical and Electronic Engineering
Imperial College London, UK
{s.mcknight18, aidan.hogg13, p.naylor}@imperial.ac.uk

***Abstract*—Evaluation of speaker segmentation and diarization normally makes use of forgiveness collars around ground truth speaker segment boundaries such that estimated speaker segment boundaries with such collars are considered completely correct. This paper shows that the popular recent approach of removing forgiveness collars from speaker diarization evaluation tools can unfairly penalize speaker diarization systems that correctly estimate speaker segment boundaries. The uncertainty in identifying the start and/or end of a particular phoneme means that the ground truth segmentation is not perfectly accurate, and even trained human listeners are unable to identify phoneme boundaries with full consistency. This research analyses the phoneme dependence of this uncertainty, and shows that it depends on (i) whether the phoneme being detected is at the start or end of an utterance and (ii) what the phoneme is, so that the use of a uniform forgiveness collar is inadequate. This analysis is expected to point the way towards more indicative and repeatable assessment of the performance of speaker diarization systems.**

***Index Terms*—Speaker diarization, forgiveness collar, phoneme boundary, diarization scoring.**

I. Introduction

A. Background

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

A widely used metric for evaluating speaker diarization systems is the diarization error rate (DER). The standard method of calculating the DER is described in detail in [4] and is based on the formula

$$DER = \frac{FA + MISS + ERROR}{TOTAL}, \tag{1}$$

where: FA is the false alarm time, which is the aggregate duration allocated to speakers when they were not speaking; $MISS$ is the missed time, which is the aggregate duration that should have been attributed to a speaker but was not; $ERROR$ is the confusion time, which is the aggregate duration attributed to the wrong speaker; and $TOTAL$ is the total speaker time, which can be longer than the speech duration because overlapping speech is attributed separately to the people who are speaking during the period of overlap.

The standard code used to evaluate speaker diarization systems is `md-eval.pl` provided by NIST as part of its scoring toolkit (SCTK) [5]. It was used for the NIST Rich Transcription challenges held from 2002 to 2009 [6] and is now also used in scoring toolkits of more recent diarization challenges such as [7]. This code enables various options to be selected, such as applying forgiveness collars of any size (historically $\pm$250 ms has been used) around ground truth segment boundaries and specifying periods that should be ignored (e.g. overlapping speech). [4] also describes other standard practices relevant to this paper, such as the recommendation that speech pauses of less than 300 ms should not be considered to separate utterances.

The rationale given in [4] for using forgiveness collars is that they account "for both the inconsistent annotation of segment times by humans and the philosophical argument of when speech begins for word-initial stop consonants". However, this use of forgiveness collars has led to DERs that can potentially give an unduly favourable impression of the diarization system, so recent diarization challenges (specifically DIHARD II in 2019 [8] and DIHARD I in 2018 [9]) have removed the forgiveness collar altogether. Those recent challenges also prohibit the exclusion of overlapping speech. Both of these requirements make sense as, intuitively, diarization systems should be evaluated on their overall performance, though as shown in this paper the removal of forgiveness collars does have possibly unintended consequences by unfairly penalizing diarization systems that make correct assumptions about utterance boundaries. That said, there are other important uses of segmentation for which forgiveness collars are still essential [10].

This paper reports an analysis of the phoneme dependency of phoneme boundary uncertainties. It is expected that this analysis would lead to more meaningful DERs and better evaluation of diarization systems. To facilitate this research, parts of the evaluation tool `md-eval.pl` have been rewritten in Python to enable the effects of different start and end collar sizes, and phoneme dependence, to be studied [11].

B. AMI Corpus and DiarTk

The AMI Corpus [12] comprises multiple recordings of 169 separate meetings with up to five speakers in each meeting, and contains detailed labelling information and transcripts. In particular for this research, 163 of the meetings

TABLE I: EFFECT OF COLLAR SIZE ON REPORTED ERROR RATES USING (a) AMI_20050204-1206 GROUND TRUTH SEGMENTS, (b) 11 MINUTE EXTRACT FROM ES2008b GROUND TRUTH SEGMENTS AND (c) REFINED AMI_20050204-1206 GROUND TRUTH SEGMENTS AS INPUT TO DIARTK BUT LESS ACCURATE ES2008b GROUND TRUTH SEGMENTS IN EVALUATION.

	(a)					(b)					(c)				
Collar (ms)	**MISS (%)**	**FALARM (%)**	**ERROR (%)**	**DER (%)**	**SBDER (%)**	**MISS (%)**	**FALARM (%)**	**ERROR (%)**	**DER (%)**	**SBDER (%)**	**MISS (%)**	**FALARM (%)**	**ERROR (%)**	**DER (%)**	**SBDER (%)**
250	2.09	0.00	6.70	8.79	40.62	4.63	0.00	6.87	11.51	46.55	13.12	0.22	6.05	19.39	54.43
200	2.93	0.00	7.30	10.23	42.47	5.90	0.00	7.18	13.07	47.88	14.16	0.24	6.37	20.78	56.36
150	4.07	0.00	8.01	12.09	44.58	7.37	0.00	7.53	14.90	49.21	15.46	0.28	6.72	22.45	58.54
100	5.39	0.00	8.89	14.29	47.12	8.86	0.00	7.96	16.82	50.98	16.86	0.32	7.16	24.34	61.36
50	6.98	0.00	9.92	16.90	50.36	10.37	0.00	8.50	18.87	53.61	18.40	0.46	7.63	26.48	65.18
0	8.74	0.35	10.94	20.03	N/A	12.04	0.14	9.08	21.26	N/A	20.15	0.91	8.09	29.14	N/A

contain complete phonemes information. The meetings left out are EN2001a, EN2001e, EN2003a, EN2006a, EN2006b and IB4005. [12] warns that the phoneme transcripts "should be used with caution", largely on the basis that [13] finds "considerable differences between human and automatic phone labelling techniques".

This research uses DiarTk [14] as an example diarization system to evaluate diarization performance because it permits separate input speaker segmentation to be used and is quite flexible. DiarTk is based on agglomerative information bottleneck principles [15], [16], which is a greedy bottom-up clustering algorithm. Although DiarTk has been chosen in this paper, this research is not limited to one specific diarization system.

C. Motivation for this Research

Table I(a) clearly shows how larger collar sizes have a significant impact on error rates. These figures are calculated using DiarTk [14] using the AMI_20050204-1206 data, specifically the ground truth speaker segments file originally provided by NIST [17] and the example Mel-frequency cepstral coefficients (MFCCs) file [18]. This example MFCCs file is derived from the full ES2008b meeting recording using 30 ms windows with 10 ms steps and uses 19 MFCCs. AMI_20050204-1206 is based on an approximately 11 minute extract (from 1,270.3 to 1,983.2 s, which is 21 mins 10.3 s to 33 mins 3.2 s) of the ES2008b meeting from the AMI corpus [12]. The last column is the segment boundary diarization error rate (SBDER), which is here defined as the errors in the utterance boundaries (i.e. instead of excluding the forgiveness collars, only the forgiveness collars are considered). All the SBDERs are much higher than the DERs, showing that diarization performance is much worse at utterance boundaries within the collars.

Table I(b) shows that using less accurate input speaker segments in diarization systems can lead to worse DERs, even where those less accurate input segments are also used as the ground truth segments in the evaluation.

Getting better results with better input segments is clearly not a surprise. However, Table I(c) shows what happens if the ES2008b speaker segments are retained as the ground truth segments file used in the NIST `md-eval.pl` evaluation code but the more refined input speaker segments are used in DiarTk. In this case, results are significantly worse than in Table I(b) for all collar sizes. This highlights how important it is to take proper account of any uncertainty in the accuracy of segmentation in the ground truth segments file used for evaluation because, otherwise, using more accurate input segments in the diarization systems actually leads to worse results in the evaluation, which is clearly not correct.

II. Evaluating Collar Importance

A. Importance of Ground Truth Labelling

To start with, it is helpful to compare specific portions of the more refined AMI_20050204-1206 speaker segments file with the corresponding 11 minute extract of the original ES2008b speaker segments provided by AMI. The AMI_20050204-1206 speaker segments file has 333 separate speaker segments across all four speakers, whereas the 11 minute extract of the ES2008b meeting has 200 speaker segments, so AMI_20050204-1206 is clearly more refined. Fig. 1 shows a 30 s extract from 1,270.3 to 1,300.3 s that is a typical illustration of the differences. In this extract, Speaker A is the main speaker but all the other three speakers say something at certain points. Note that all the "sp" phonemes that denote short pauses are so short that in Fig. 1 their boundaries appear to immediately lead into the next phoneme.

A number of points are evident from Fig. 1. In particular, the main sentence uttered by Speaker A ends in "*before the meeting and that we have to keep in in um in mind as we're creating this*.", but there are several significant pauses that pose problems for speaker diarization systems. One of these occurs after "*meeting*", determined by AMI_20050204-1206 to be 315 ms long and breaking the utterance, but the original ES2008b labelling does not detect a break. Another occurs after "*keep in*", which neither system determines is significant enough to warrant breaking the utterance, though the phonemes register a "sil" (i.e. a silence) of 180 ms, and eventually Speaker A stutters slightly before resuming.

References to "start" or "starting" phoneme mean the phoneme identified as being the first phoneme in an utterance, and similarly "end" or "ending" phoneme means the phoneme identified as being the last phoneme in an utterance.

B. Assessing Phoneme Relevance

Fig. 2 shows three histograms: the first is all phonemes used in 163 AMI Corpus meetings; the second is the start phonemes; and the third is the end phonemes. There are 46 phonemes used altogether in the AMI Corpus, but for consistency with other research this paper follows the reduced phonemes methodology in [19] to reduce to 38 phonemes. The phonemes reduced are "ao" $\rightarrow$ "aa", "ax" $\rightarrow$ "ah", "axr" $\rightarrow$ "er", "zh" $\rightarrow$ "sh", "em" $\rightarrow$ "m", "en" $\rightarrow$ "n", "el" $\rightarrow$ "l" and "sp" $\rightarrow$ "sil", though in some cases in this research

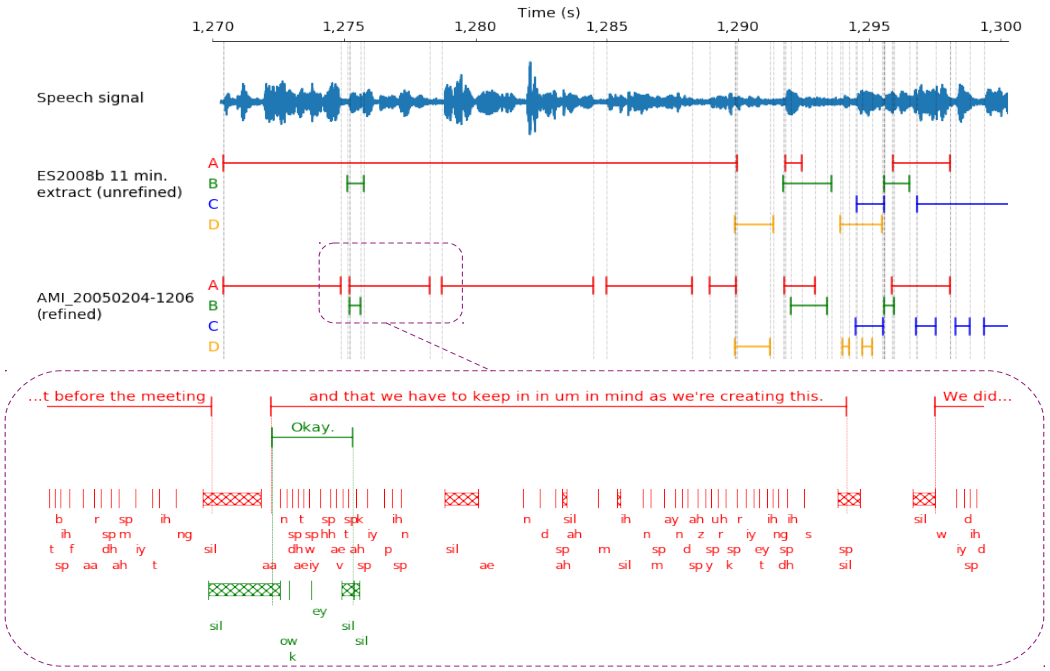

Fig. 1: 30 s extract highlighting important differences between the more refined AMI_20050204-1206 speaker segments and the less refined AMI ES2008b speaker segments for ES2008b

"sp" is retained to illustrate the differences between a short pause and a longer pause that may signify the start or end of an utterance. Note that although [19] reduces the original 61 phonemes identified in the TIMIT dataset to 39, the extra phoneme is "dx" which never appears at the start or at the end of an utterance so is not relevant for this research.

The phonemes most relevant for this research are those at the starts and ends of utterances. Fig. 2(b) and 2(c) show the histograms of the start and end phonemes for all utterances located in 163 AMI Corpus meetings. The methodology used to locate these phonemes is the same as described in Section II-D up to the widest collar of ± 1 s, so start and end phonemes were not found for all utterances.

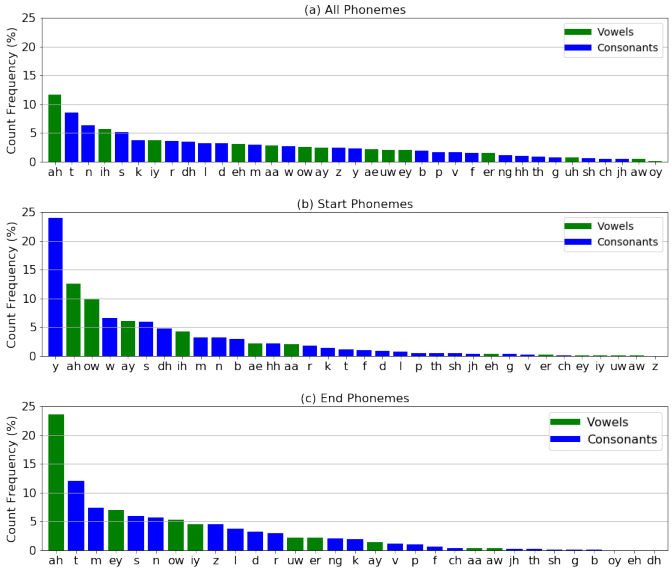

Fig. 2: Histogram of (a) all phonemes (b) start phonemes and (c) end phonemes, each in 163 AMI Corpus meetings

Of the 37 possible utterance starting phonemes, 34 occur at the start of an utterance and 31 at the end of an utterance. In both cases, some phonemes occur far more frequently than the others, but there are considerable differences between the phonemes that occur at the start of utterances compared to the end. For example, "y" is by far the most common starting phoneme, but it never appears at the end of an utterance. By contrast, "ah" is the second most common starting phoneme but is also the most common ending phoneme.

C. Phoneme Duration Uncertainty

Phonemes are known to have widely varying durations [20]. In the AMI Corpus, the average duration of all non-silence phonemes is 87.46 ms, which is considerably shorter than both the 118.35 ms average starting phoneme length and the 282.06 ms average ending phoneme length. There is considerable variation in the mean durations of individual phonemes and their standard deviations, as highlighted in Fig. 3 along with the maximum and minimum durations of each phoneme.

The phoneme duration uncertainty means that typical phoneme durations probably cannot be used to help identify when a particular phoneme starts or ends. Consequently, this research analyses utterance boundary uncertainties instead.

D. Locating Utterance Start and End Phonemes

It is informative to look at the ground truth speaker segments of the AMI Corpus and the phonemes identified as the starting and ending phonemes at the utterance boundaries. Starting from the ground truth speaker segments file, the phonemes at the start and end of each utterance were analysed by starting with a small 5 ms collar around the utterance boundary time. Then, with the speaker identified by the ground truth speaker

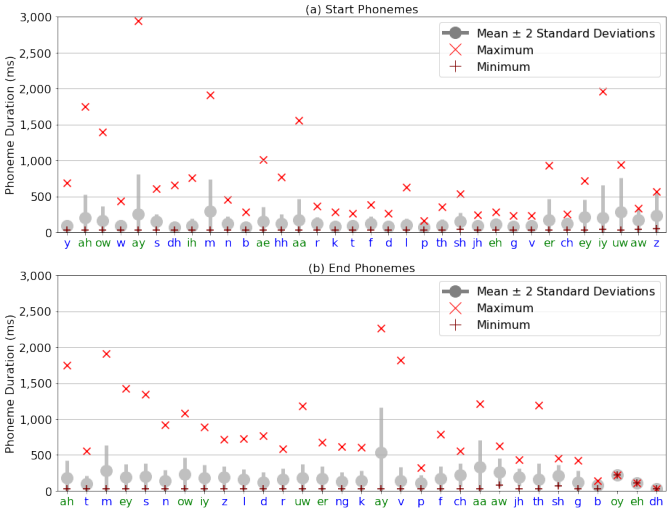

Fig. 3: Start and end phoneme duration means, standard deviations, maximum and minimum values

segments file, for each speaker start utterance any "sil" then phoneme boundaries in that collar spoken by the relevant speaker were identified. If there was only one such phoneme boundary, it was determined to be the utterance start. If there was more than one, the "sil"/phoneme boundary nearest to the ground truth start time was determined to be the start. A similar analysis was carried out for utterance ending phonemes by locating phoneme then "sil" boundaries in the collar. All utterances are assumed to comprise phonemes between a starting "sil" and an ending "sil".

Fig. 4 shows the number of start and end phonemes identified for individual meetings within a particular collar size, expressed as a percentage of the total number of utterances of that meeting, along with averages over all meetings.

None of the meetings score 100% start or end phoneme identification accuracy. The best results with a collar of ± 1 s are 91.23% start phonemes and 90.44% end phonemes both for TS3004c, and the worst are 22.05% start phonemes for TS3004b and 20.06% end phonemes for IB4002. However,

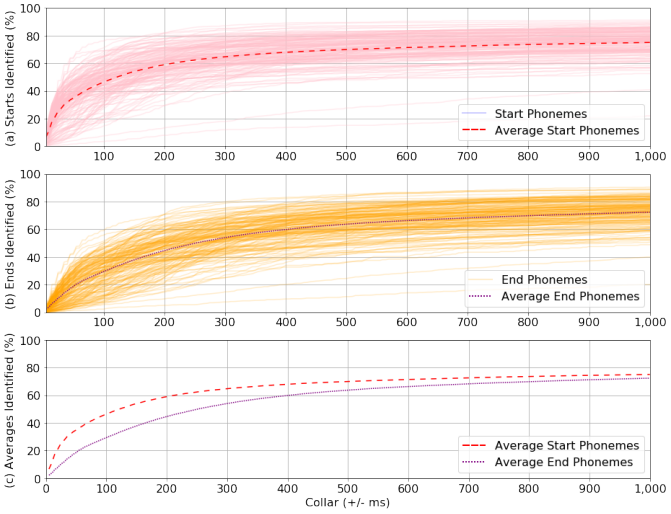

Fig. 4: Percentage of phonemes identified at utterance start/end boundaries with increasing collar sizes for 163 AMI Corpus meetings

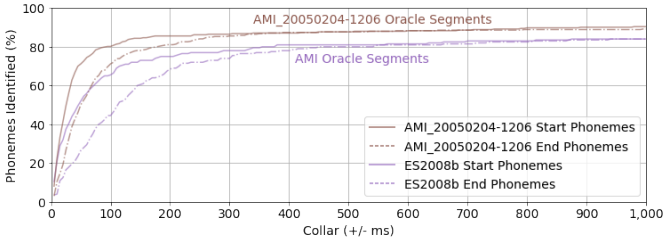

Fig. 5: Percentage of starting/ending phonemes found for AMI_20050204-1206 v. ES2008b ground truth segments

there are two particularly bad outlier meetings - IB4002 and EN2002a, where fewer than 50% of start and end phonemes are identified with a collar of ± 1 s.

Identifying utterance start phonemes generally performs better than identifying utterance end phonemes, so forgiveness collars should be wider when applied at the end of utterances.

Fig. 5 compares the identification of utterance start and end phonemes of AMI_20050204-1206 and the 11 minute extract of ES2008b. The hit rate is considerably better for AMI_20050204-1206, reaching 90.4% for start phonemes and 89.2% for end phonemes, compared to 84.0% and 84% for the 11 minute extract of ES2008b. This seems counterintuitive as ES2008b has fewer speaker segments so it would be expected to be easier to identify the phonemes at the starts and ends, but the more refined segmentation of AMI_20050204-1206 tallies better with the phoneme boundaries.

E. Start v. End of Utterance Uncertainty

Quantifying the uncertainty in the starting and ending phonemes is challenging. One method used here calculates the means for each phoneme of the distance between the ground truth segment start time u_{gt} and the start time of the relevant phoneme u_j. Let p_i denote a possible start or end phoneme, so the set $P = \{p_1, p_2, ..., p_{37}\}$ denotes all 37 such phonemes. The starting means $\mu_{u,i}$ for each phoneme p_i are calculated by iterating over each start phoneme $s_j \in S$, and if s_j equals p_i evaluating the distance. The set of starting phonemes $S = \{s_1, s_2, ..., s_N\}$ and N = number of utterances. Each starting phoneme s_j will be matched to a phoneme p_i if one is located within the relevant collar, though not all are. For each phoneme p_i, the starting means are

$$\mu_{u,i} = \frac{\sum_{j=1}^{N} \delta_{i,j}\left(u_j - u_{gt}\right)}{\sum_{j=1}^{N} \delta_{i,j}}, \quad \delta_{i,j} = \begin{cases} 1 & \text{if } s_j = p_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and the unbiased starting standard deviations $\sigma_{u,i}$ from

$$\sigma_{u,i}^2 = \frac{\sum_{j=1}^{N} \delta_{i,j}\left(u_j - \mu_{u,i}\right)^2}{\left(\sum_{j=1}^{N} \delta_{i,j}\right) - 1}. \tag{3}$$

The ending means and standard deviations for each phoneme p_i are calculated similarly.

Fig. 6(a) and 6(b) show the means and standard deviations for the starting phonemes, and Fig. 6(c) and 6(d) show the

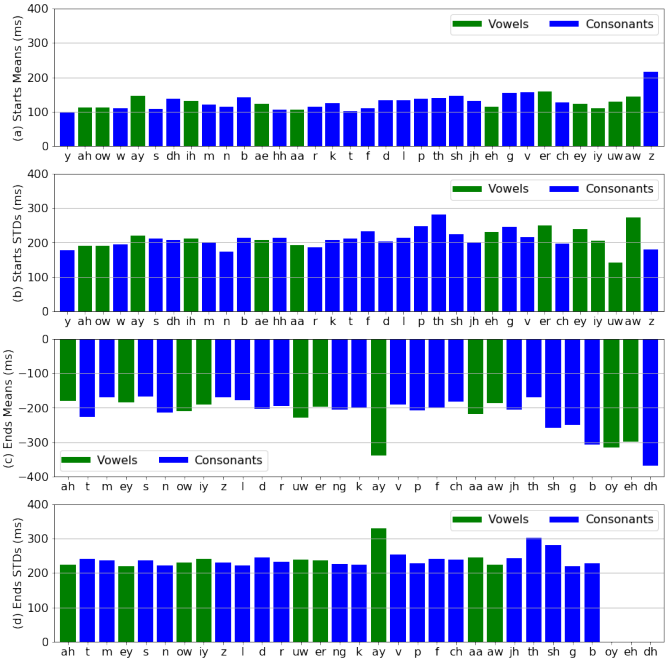

Fig. 6: Start and end phoneme mean distances and their standard deviations (STDs) from the ground truth segments (ordered by decreasing frequency)

means and standard deviations for the ending phonemes. The AMI Corpus ground truth speaker segments clearly provide for significantly longer utterances than the starting and ending phonemes would suggest as the mean distances are all positive for the starting phonemes and all negative for the ending phonemes. In Section II-C it was shown that the starting and ending phonemes already have considerably longer mean durations than other phonemes, yet the longer ground truth segments suggests that even they are an underestimate.

Most utterances tend to taper out [21], [22], so longer end standard deviations are not surprising and should be factored into evaluation tools.

There is some variation across phonemes, which could be exploited by evaluation tools. For example, the "y" start phoneme should not need a large forgiveness collar, but the "ay" end phoneme would. A system could be set up whereby the utterance start and end times are calculated in a manner consistent with the phoneme detection, so the means in Fig.6 should tend to zero and leave narrow standard deviations that could be used meaningfully by an evaluation tool.

III. Conclusion

This paper shows that there is considerable uncertainty in determining exact start and end times of utterances, which can lead to inaccuracies in ground truth segmentation that unfairly penalize speaker diarization systems that correctly determine when utterances should start and end. This research suggests that the popular recent approach of removing forgiveness collars altogether from diarization challenges is not ideal for typical labelling. Evaluation tools that account for phonemes at utterance boundaries and whether they appear at the start or at the end of an utterance could give a better assessment of the performance of diarization systems.

References

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

[2] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, Sept. 2018, pp. 2808–2812.

[3] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Černocký, "Bayesian HMM Based x-Vector Clustering for Speaker Diarization," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, Sept. 2019, pp. 346–350.

[4] NIST, "The 2009 (RT-09) rich transcription meeting recognition evaluation plan," Feb. 2009.

[5] ——, `md-eval.pl` (Version 22), in `sctk-2.4.10`, [Online]. Available: ftp://jaguar.ncsl.nist.gov/pub/sctk-2.4.10-20151007-1312.tar.bz2.

[6] ——, "Rich transcription evaluation", [Online]. https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation.

[7] N. Ryant, "dscore: Diarization scoring toolkit", 2019, [Online]. Available: https://github.com/nryant/dscore.

[8] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "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.

[9] ——, "First DIHARD challenge evaluation plan, Tech. Rep.," 2018. [Online]. Available: https://zenodo.org/record/1199638#.XkABaWj7Q2w

[10] A. O. T. Hogg, C. Evers, and P. A. Naylor, "Multiple hypothesis tracking for overlapping speaker segmentation," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 2019, pp. 195–199.

[11] S. W. McKnight, A. O. T. Hogg, and P. A. Naylor, "pyDERCalc," 2020. [Online]. Available: https://github.com/swm1718/pyDERCalc

[12] J. Carletta, S. Ashby, S. Bourban, M. Flynn, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, and M. W. P. D. Reidsma, "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.

[13] S. Greenberg, "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2-4, pp. 159–176, Nov. 1999.

[14] D. Vijayasenan and F. Valente, "DiarTk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2012, pp. 2170–2173.

[15] D. Vijayasenan, F. Valente, and H. Bourlard, "Combination of agglomerative and sequential clustering for speaker diarization," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2008, pp. 4361–4364.

[16] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Proc. Neural Inform. Process. Conf*, S. A. Solla, T. K. Leen, and K. Müller, Eds., 2000, pp. 617–623.

[17] IDIAP, `AMI_20050204-1206.rttm`, [Online]. Available: https://github.com/idiap/IBDiarization/tree/master/data/rttm.

[18] ——, `AMI_20050204-1206.fea`, [Online]. Available: https://github.com/idiap/IBDiarization/tree/master/data/mfcc.

[19] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[20] K. Silverman and J. Bellegarda, "Using a sigmoid transformation for improved modeling of phoneme duration," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 1, Mar. 1999, pp. 385–388.

[21] A. Cutler and M. Pearson, "On the analysis of prosodic turn-taking cues," *Intonation Discourse*, pp. 139–156, 1986.

[22] L. Ferrer, E. Shriberg, and A. Stolcke, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 1, Apr. 2003, pp. 608–611.