

Where did I go wrong?: Identifying troublesome segments for speaker diarization systems

Mary Tai Knox^{1,2}, Nikki Mirghafori¹, Gerald Friedland¹

¹International Computer Science Institute, Berkeley, California, USA

²Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA

{knoxm, nikki, fractor}@icsi.berkeley.edu

Abstract

The focus of this work is to identify types of segments that are difficult for speaker diarization systems. The diarization outputs of five state-of-the-art systems are analyzed on short/long segments as well as segments surrounding speaker change-points. We found that for all five systems as the duration of the segment decreased the diarization error rate (DER) increased. Also, segments immediately preceding and following speaker change-points performed much worse than their respective counterparts. In fact, at least 40% of the DER for all five systems is attributed to time within 0.5 seconds of a speaker change-point. We hope the results of this work motivate future improvements of speaker diarization systems.

Index Terms: speaker diarization, error analysis, rich transcription

1. Introduction

The goal of speaker diarization is to partition an audio signal into speaker homogeneous speech regions, as shown in Figure 1, where the number of speakers as well as the speaker identities are not known a priori. Speaker diarization has many applications, including speaker adaption for automatic speech recognition and audio indexing.

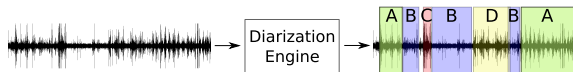


Figure 1: Overview of speaker diarization. From an input audio signal, segment the signal into nonspeech and speech segments, the latter labeled by speaker (e.g., A, B, C, D).

Most of the effort in speaker diarization research is spent introducing new and improved algorithms to “solve” the speaker diarization problem, and relatively less work is on analyzing speaker diarization performance and error patterns. The focus of this paper is to analyze speaker diarization errors for the meeting domain. More specifically, we investigate speaker diarization performance for five state-of-the-art speaker diarization systems on specific types of segments (e.g., short/long segments and before/after speaker changes).

Previous analyses of speaker diarization have focused on two main methods. The first compares performance between systems using characteristics of the meeting, such as the number of speakers and average conversation turn duration [1, 2]. In [1], Mirghafori and Wooters studied how characteristics of entire broadcast news recordings correlated with Diarization Error

Rate (DER) derived statistics. In [2], Bozonnet et al. noted that the top-down system better estimated the true number of speakers while the bottom-up system output typically had a more similar number of segments and average segment duration to the speaker diarization reference.

The second method of performing speaker diarization error analysis involves replacing components of a given system with oracle components and calculating the effect on the DER [3, 4]. For example, Huijbregts et al. began with an oracle diarization system, where each component (e.g., speech activity detection, initialization, merging criterion, stopping criterion) was an oracle component which utilized the reference transcription. In both a top-down and bottom-up fashion, each oracle component was replaced with its speaker diarization system component and the change in DER before and after the replacement reflected the effect that component (and potentially subsequent components) had on the DER.

The previous analysis work focused on correlating diarization performance with attributes of the recording and computing the change in DER associated with each component of the system. By contrast, the goal in this work is to characterize which types of segments are difficult for five state-of-the-art speaker diarization systems. The results of this work provide insight into where speaker diarization researchers should focus their attention in order to further improve speaker diarization as well as understand the pitfalls of the various speaker diarization algorithms.

This paper is outlined as follows: in Section 2 we describe the experimental setup used in this analysis, in Section 3 we define the types of segments investigated, in Section 4 we provide and discuss the results, and in Section 5 we give our conclusions as well as areas of future work.

2. Experimental Setup

2.1. Data

This analysis is performed on the NIST Rich Transcription '09 (RT-09) evaluation dataset. The RT-09 dataset consists of seven meetings recorded at three sites: IDIAP, Edinburgh, and NIST. Both the multiple distant microphone (MDM) and single distant microphone (SDM) conditions are investigated. However, due to space constraints, only the MDM results are presented in this paper. Note that the same trends seen for the MDM condition also occurred for the SDM condition.

2.2. Speaker Diarization Systems

The output segmentation files from five speaker diarization systems (AMI [5], ICSI [6], IDIAP [7], IIR-NTU [8], and LIA-

Eurecom [9]) are analyzed in this work. These systems represent the state-of-the-art in speaker diarization and have consistently performed well in the NIST Rich Transcription evaluations. The system results are anonymized since our aim is not to identify the best performing system but instead to identify trends among systems.

2.3. Scoring Metric

The Diarization Error Rate (DER) defined by NIST [10] is used to evaluate each system’s performance. In order to compute the DER, first an optimal one-to-one mapping of reference speakers to system output speakers is determined. The DER is then the sum of the per speaker false alarm time (overestimating the number of speakers), miss time (underestimating the number of speakers), and speaker error time (the hypothesized speaker(s) is (are) not matched to the appropriate speaker(s) in the reference) divided by the total speech time in an audio file, as shown in Equation (1). As done in the NIST evaluations, we scored the DER using a *no-score collar* of ± 0.25 seconds [11] around reference segment boundaries. Overlapping speech errors have been a long-standing, known source of speaker diarization error [12, 13]. Therefore, each of the three types of errors (T_{FA} , T_{MISS} , and T_{SPKR}) are further split into times during overlapping speech and during single speaker speech.

$$DER = \frac{T_{FA} + T_{MISS} + T_{SPKR}}{T_{SPEECH}} \quad (1)$$

Note that each of the segment types studied in this work are labeled using the reference transcription. Therefore, nonspeech time (as transcribed in the reference) is not scored in this study. Thus, the only way to have a false alarm error would be if a system is able to hypothesize overlapping speech, or more than one speaker speaking at a given instance. Only one of the five systems hypothesizes overlapping speech. The other four systems annotate at most one speaker at a given time. In order to retain the anonymity of the systems, the false alarm errors are not shown. We feel that this does not affect the results of this paper since the false alarm error rate during speech time is negligible.

3. Segment Types

Speaker diarization performance is evaluated for two types of segments: segments categorized based on the segment duration and segments categorized based on their proximity to speaker changepoints. In this work, a *segment* is defined according to the reference speaker diarization segmentation. The reference segmentation is created by first force aligning the individual headset microphone audio to the reference transcripts using LIMS tools. Then the word boundaries obtained from the forced alignment are smoothed using a 0.3 second window, thereby grouping multiple words together into a segment [11].

3.1. Segment Duration

Speaker diarization system performance is evaluated based on the duration of each segment. More specifically, the DER is computed for 10 bins of segment durations, where each bin contains segments of similar duration. For illustration, Figure 2 shows an example reference segmentation which is split into three bins (short, intermediate, and long segments).

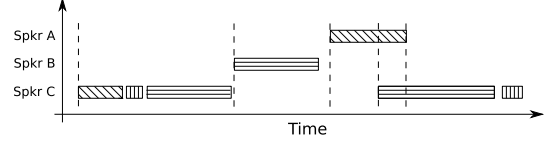


Figure 2: Example reference segmentation containing three speakers (A, B, and C). Changepoints are represented using vertical dashed lines. The short, intermediate, and long segments are filled with vertical, diagonal, and horizontal lines, respectively.

3.2. Speaker Changepoints

Segments surrounding speaker changepoints are also examined. In this work, a *speaker changepoint* is defined as an instance in which the current speaker(s) differs from the previous speaker(s). Nonspeech segments are ignored since most speaker diarization systems similarly discard these segments [5, 6, 7, 8, 9]. Thus, if a speaker talks for some time, pauses, and then resumes talking there is no speaker changepoint when the speaker resumes talking. A segment is labeled a first segment after a speaker changepoint (*FirstAfter*) if any portion of the segment immediately follows a speaker changepoint. Similarly, a segment is labeled a last segment before a speaker changepoint (*LastBefore*) if any portion of the segment is contained in the last segment prior to a changepoint. Examples of FirstAfter and LastBefore segments are shown in Figure 3.

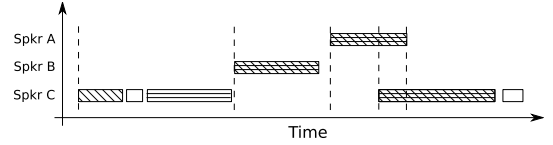


Figure 3: Visualization of FirstAfter and LastBefore segments. Segments filled with diagonal lines are FirstAfter segments. Segments filled with horizontal lines are LastBefore segments.

4. Results

The DERs for each of the systems are computed over segments of similar duration. The scored segments are split into 10 bins, each containing a roughly equivalent number of segments. For each bin, the DER is calculated for all five systems. The results are shown in Figure 4. All five systems display the same trend: as the duration of the reference segment increases, the DER improves. Both the miss rate and the speaker error rate decrease as the duration of the segments increase, though the miss rate (particularly due to overlapping speech) plays a larger role in the decreasing DER. Note that due to the ± 0.25 second collar, the minimum scored segment duration time is 0.51 seconds. Also, for the following plots the DER is color coded according to the type of error it is (miss and speaker error). The miss and speaker error rates are further split into times containing overlapping and single speaker speech. The miss rates during overlapping and single speaker speech are annotated as red and light red, respectively. Similarly, the speaker error rates during overlapping and single speaker speech are annotated as blue and light blue, respectively.

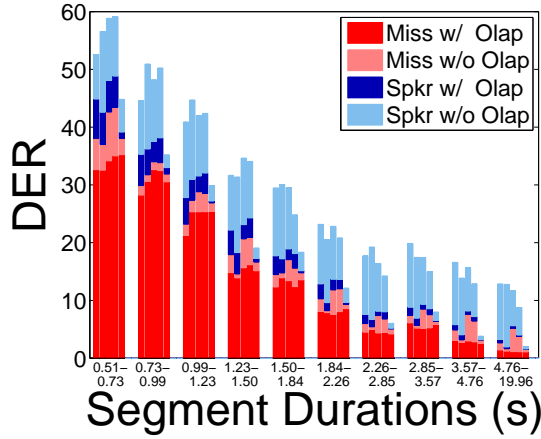


Figure 4: DERs for various segment duration bins. Each bin contains roughly the same number of segments. Each bar represents the DER for one of the five analyzed systems. Spkr denotes speaker error, olap denotes overlapping speech.

The errors surrounding speaker changepoints are also examined. In Figure 5, the DERs for each of the systems are shown for segments following a changepoint (FirstAfter) and the complement. FirstAfter segments perform significantly worse than not FirstAfter segments both in terms of miss rate (particularly due to misses occurring during overlapping speech) and speaker error rate. Similar results are obtained for the segments immediately before a changepoint (LastBefore) as shown in Figure 6. Since a segment is classified as FirstAfter if *any* portion of the segment immediately follows a changepoint (and similarly for LastBefore segments), not FirstAfter and not LastBefore segments did not contain overlapping speech.

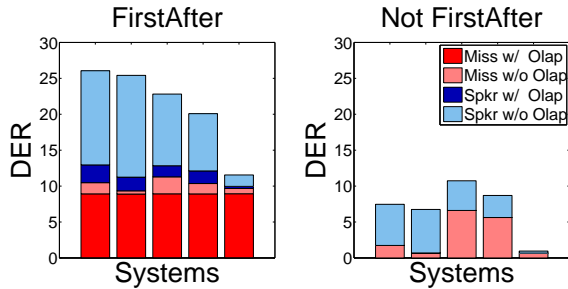


Figure 5: DERs for segments following and not following changepoints.

Both short segments and segments preceding or following speaker changepoints performed worse than their counterparts. In order to verify that these are in fact two separate types of errors (and it is not the case that segments preceding and following speaker changepoints are dominated by short segments) we computed the cumulative distribution functions (CDFs) of the segment durations for segments immediately after and preceding speaker changepoints as well as their respective complements. The distributions are shown in Figure 7. The CDFs of the segment durations for FirstAfter and LastBefore segments lie on top of one another. Similarly, the CDFs for not FirstAfter and not LastBefore segments overlap. In fact, all four CDFs are

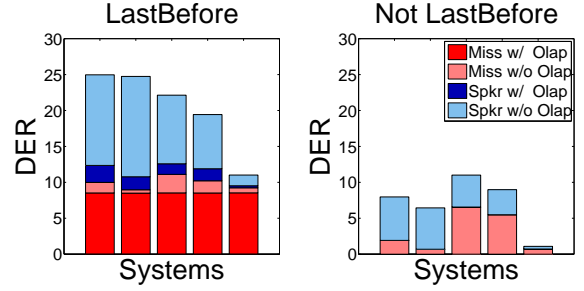


Figure 6: DERs for segments preceding a changepoint and not preceding a changepoint.

quite close. Thus, FirstAfter and LastBefore segments do not contain an unusually high number of short segments.

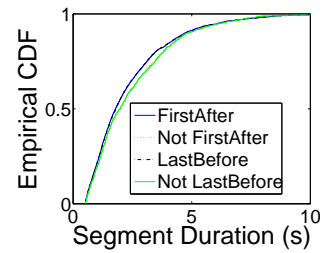


Figure 7: CDFs of segment durations for FirstAfter, not FirstAfter, LastBefore, and not LastBefore segments. Note that the CDFs are close to one another.

Next, we investigate the time surrounding speaker changepoints in more detail. Instead of grouping an entire segment together, we split each segment into 0.25 second intervals. We then plot the DER as a function of the time after/until the previous/next changepoint as shown in Figure 8. This figure demonstrates that the systems have a more difficult time closer to changepoints than farther away from changepoints. Once again both the miss and speaker error rates decrease as the time from the changepoint increases, with the miss rate contributing more to the dramatic decrease in DER. Note that the last bin combines all of the time greater than 2.5 seconds from the changepoint.

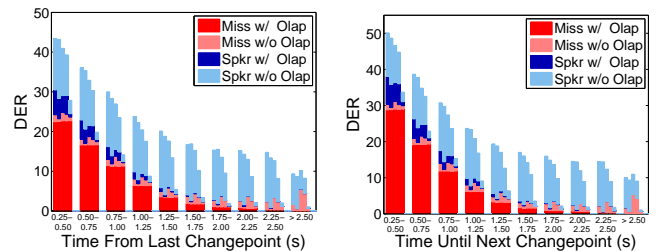


Figure 8: DER as a function of the time since the last speaker changepoint and until the next speaker changepoint for all five systems.

We then analyze the results when measuring the time to the closest changepoint, regardless of whether the changepoint is before or after the given instance. The results are shown in

Figure 9. In this case, the DER initially decreases dramatically and then remains for the most part steady. Then in Figure 10, we show the percent of scored time associated with each of the ten distances from the changepoint, as defined on the x-axis of Figure 9, on the left. On the right is the percent of each system's DER for each of the distances from the changepoint. For all five systems, at least 40% of the DER occurred between 0.25 and 0.50 seconds from the speaker changepoint.

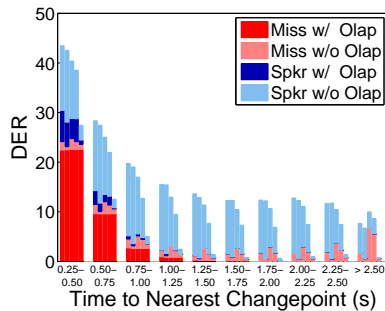


Figure 9: DER as a function of the distance to the closest changepoint, which could be before or after the given instance.

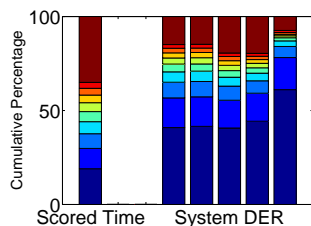


Figure 10: Percent of scored time contained in each distance from the changepoint (as defined in the previous figure) and percent of system DER contained in each distance from the changepoint.

5. Conclusions and Future Work

In conclusion, we have demonstrated two problematic types of segments for speaker diarization systems. Both short segments and segments surrounding speaker changepoints caused a considerable amount of DER for all five state-of-the-art speaker diarization systems. We have further noted that DER increases closer to the changepoint, with at least 40% of the DER occurring within 0.50 seconds of the changepoint.

At initial inspection, it may not be surprising that performance degrades for short segments and near speaker changepoints. Some systems utilize a minimum duration constraint [6], not allowing for very short segments. With regard to the difficulty of segments surrounding speaker changes, it is important to note that all of the diarization systems used in this study are offline, so it is interesting to see that performance degrades both before and after speaker changes.

In the future, we plan to investigate these segments further to determine the causes of this poor performance. We hypothesize that the causes may not be solely due to limitations of the systems but perhaps a result of speakers not using speaker discriminative words or speakers modifying speech patterns to take

the floor or allow the floor to be taken. Another area of future work is to explore how the differences in speaker diarization algorithms affect the DER, both overall and for various types of segments. We hope that shedding light on the significance of these particular sources of error paves the way for development of targeted strategies to overcome them.

6. Acknowledgements

We would like to thank Marijn Huijbregts (AMI), Haizhou Li (IIR-NTU), Fabio Valente (IDIAP), Deepu Vijayasenan (IDIAP), and Simon Bozonnet (LIA-Eurecom) for generously sharing their speaker diarization outputs. Without their support, this paper would not have been possible. This work is partially supported by the NSF under grant OISE-1135365.

7. References

- [1] N. Mirghafori and C. Wooters, "Nuts and flakes: A study of data characteristics in speaker diarization," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. 1017–1020.
- [2] S. Bozonnet, N. Evans, X. Anguera, O. Vinyals, G. Friedland, and C. Fredouille, "System output combination for improved speaker diarization," in *Interspeech*, Makuhari, Japan, 2010.
- [3] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *Interspeech*, Antwerp, Belgium, August 2007.
- [4] M. Huijbregts, D. van Leeuwen, and C. Wooters, "Speaker diarization error analysis using oracle components," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 393–403, February 2012.
- [5] M. Huijbregts, "Segmentation, diarization and speech transcription: surprise data unraveled," Ph.D. dissertation, University of Twente, Enschede, Netherlands, November 2008.
- [6] G. Friedland, A. Janin, D. Imseng, X. Anguera Miro, L. Gottlieb, M. Huijbregts, M. Knox, and O. Vinyals, "The ICSI RT-09 speaker diarization system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 371–381, February 2012.
- [7] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream speaker diarization beyond two acoustic feature streams," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, March 2010, pp. 4950–4953.
- [8] T. Nguyen, H. Sun, S. Zhao, S. Khine, H. Tran, T. Ma, B. Ma, E. Chng, and H. Li, "The IIR-NTU speaker diarization systems for RT'09," in *RT'09, NIST Rich Transcription Workshop*, Melbourne, Florida, 2009.
- [9] S. Bozonnet, N. Evans, and C. Fredouille, "The LIA-Eurecom RT'09 speaker diarization system: Enhancements in speaker modelling and cluster purification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, March 2010.
- [10] NIST, "The 2009 (RT-09) rich transcription meeting recognition evaluation plan," 2009.
- [11] J. Ajot and J. Fiscus, "RT-09 speaker diarization results," Melbourne, Florida, May 2009.
- [12] S. Otterson, "Use of speaker location features in meeting diarization," Ph.D. dissertation, University of Washington, Seattle, Washington, 2008.
- [13] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved diarization in multi-party party meetings," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, 2008, pp. 4353–4356.