# CHARACTERIZING PERFORMANCE OF SPEAKER DIARIZATION SYSTEMS ON FAR-FIELD SPEECH USING STANDARD METHODS

*Matthew Maciejewski*[1] *, David Snyder*[1,2] *, Vimal Manohar*[1,2] *, Najim Dehak*[1] *, Sanjeev Khudanpur*[1,2]

[1] Center for Language and Speech Processing
[2] Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA

## ABSTRACT

To date, the bulk of research on speaker diarization has been conducted on telephone or near-field speech. As the need for technologies capable of handling conversational speech increases, it is necessary to establish the performance of state-of-the-art systems in this domain. In this work we evaluate the performance of an ivector/PLDA-based diarization system on the AMI Meeting Corpus, comparing performance on near-field, far-field, and signal-enhanced conditions.

***Index Terms***— Speaker diarization, far-field speech

## 1. INTRODUCTION

Put simply, speaker diarization is answering the question: "Who spoke when?" More thoroughly, diarization is the task of identifying where speech is in a recording, segmenting it, and labeling each segment according to its speaker. As such, diarization is a critical part of any application dealing with speech with multiple speakers present.

Recently, the bulk of work on diarization has been restricted to telephone speech, which is near-field, narrow-band, and typically contains only two speakers [1, 2, 3]. While some attention has been given to broadcast news [4], nevertheless this is not representative of many desired applications. For example, voice-controlled home assistants use far-field microphones and can face any number of people speaking. In addition, health and behavioral researchers would like to be able to automatically process interactions that have been captured with a fixed microphone. It has been shown that automatic speech recognition systems in these types of situations perform worse than in the near-field case and require special treatment to improve performance, and it is unlikely that this would not be the same for diarization systems.

In this work, we demonstrate variations in diarization performance across multiple conditions and establish baselines using standard state-of-the-art diarization systems using the AMI Meeting Corpus [5]. This corpus consists of meetings of 4-5 people recorded simultaneously over a number of microphones, particularly a tabletop array, making it uniquely valuable in establishing diarization performance in conditions similar to current applications.

## 2. DATA

Our experiments were conducted using the AMI Meeting Corpus, which consists of roughly 100 hours of recorded meetings which have been manually transcribed. Each meeting consists of 4-5 participants who were recorded while roleplaying a design team, though roughly one-third of the recordings are spontaneous meetings. The meetings were recorded simultaneously over a number of microphones. We used audio recordings from two sources: head-mounted microphones and tabletop array distance microphones.

For our experiments, we constructed three conditions for diarization. The first is a single microphone from the tabletop array (referred to as 'sdm1'). The second is a beamformed [6] combination of the eight microphones in the tabletop array (referred to as 'mdm8'). The final is a summation of the headset microphones to form a synthetic near-field diarization condition (referred to as 'hms'). The corpus was split according to the standard AMI ASR partition[1] into 'dev' and 'eval' sets, each of approximately 10% of the full corpus, which together make the test set, leaving the remaining 80% reserved for training.

An important facet of the corpus is that it is very small, with a total of roughly 200 speakers and 100 hours of speech. One effect of this is that the test set of the combined 'dev' and 'eval' splits is small enough that the error rate can be noisy. In addition, in cases where external training data is not used, care must be taken to efficiently use the limited training data.

Ground truth for the training and scoring was generated automatically from the time-aligned manual transcripts of the head-mounted microphones from the corpus release. It is worth noting that in some recordings the corpus-collection proctor's speech is present in the recordings, but since he or she did not have a head mic, their speech was not transcribed. We chose to omit speech activity detection from our pipeline and used oracle segmentation, but systems using speech ac-

[1] http://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml

tivity detection will likely see slightly inflated error rates due to incorrect speech detection false alarms for regions where the constructed ground truth is missing the proctor's speech.

## 3. EVALUATION METRIC

Our choice of evaluation metric is Diarization Error Rate (DER), which is a relatively standard scoring method for speaker diarization, as developed by NIST. DER is effectively the fraction of time that has been labelled with the incorrect speaker or as non-speech. In our experiments we choose to use oracle speech activity marks, so the only errors come from misattribution of speech to an incorrect speaker. In addition, the standard parameters on the scoring script omit from scoring both regions with overlapping speech from multiple speakers and regions within a collar around speaker boundaries.

## 4. METHODS

### 4.1. General Diarization System

Typical diarization systems consist of four main components: speech activity detection, speaker representation, clustering, and resegmentation [7]. The first part is speech activity detection (SAD), i.e. labeling the regions where speech is present. This can also include change-point detection for initial segmentation into unlabeled speaker turns. The next step is to extract some kind of speaker representation from the speech. This is done either over a fixed-length sliding window or over speaker turns. The next step is to perform clustering of the segments, using a score or distance metric. Finally, there is often some kind of resegmentation, using the learned speaker labels to readjust the initial speaker boundaries.

For our experiments, we chose to use ivectors [8] for the speaker representation with a PLDA [9] for scoring and agglomerative hierarchical clustering [10], omitting speech activity detection and resegmentation. Our system was built using the Kaldi speech recognition toolkit [11].

We chose to use oracle speech activity labels over a SAD system for a number of reasons. While we acknowledge that speech activity detection is a challenging problem, particularly in the far-field case, the focus of our work was on the core of the diarization system, and the comparison across conditions is more significant with identical segmentation. In addition, the DER metric can be somewhat noisy with respect to variations in the speech activity labelling, particularly at higher error rates.

We chose to omit resegmentation from our system due to generally relatively minor gains. This is in part because the standard scoring options for DER include a generous collar of 0.5 total seconds around each speaker boundary that is omitted from scoring.

Details on the specifics of the components of the system are as follows.

#### *4.1.1. Ivector*

For the speaker representation part of the diarization system we use a relatively standard ivector system to extract ivectors over a short sliding window over speech regions, leaving the speaker boundaries to be discovered through the clustering algorithm. We chose a window size of 1.5 seconds with a window shift of 0.75 seconds. We chose this approach over using change-point detection to discover speaker segments and extracting ivectors over those segments due to the belief that giving each ivector a constant number of frames is more valuable than getting potentially longer, more stable ivectors, particularly since turns can end up being quite short a significant portion of the time.

The ivector extractor itself is close to a standard ivector extractor from a speaker recognition system, though smaller and simpler due to the less strict requirements on speaker labels of a diarization system compared to a speaker identification system. We use MFCCs for wide-band 16kHz audio with first-order deltas as input features. The ivector extractor uses 2048 gaussians with a final ivector dimension of 128.

#### *4.1.2. PLDA*

We use Probabilistic Linear Discriminant Analysis (PLDA) [9] to produce pairwise scores between the ivectors. It is trained on ivectors extracted from the training data over a window of 3 seconds to better match the runtime ivectors. Before being passed to train the PLDA, the ivectors have the global mean subtracted, are passed through a whitening transform, and are length-normalized.

#### *4.1.3. Clustering*

For clustering, we use Agglomerative Hierarchical Clustering (AHC) using pairwise PLDA scores between all of the ivectors in a test recording. This is a "bottom-up" clustering approach where each ivector is assigned to a cluster and each cluster is merged according to the PLDA score until a stopping criterion is met.

The stopping criterion is computed using unsupervised calibration of the PLDA scoring [12]. K-means clustering is used to fit two clusters to the test condition PLDA scores, with the average of the centroids as the stopping threshold. The 'dev' and 'eval' sets are used as held-out sets for each others' stopping criterion.

### 4.2. Modifications for AMI Corpus

#### *4.2.1. Speaker Representation Variation*

We explored two alternatives for the 'speaker representation' part of the diarization pipeline. The first was to, rather than

| condition | baseline | speaker multiplier | | | | | | segment-based |
|---|---|---|---|---|---|---|---|---|
| | | 5 | 10 | 50 | 100 | 150 | 200 | |
| sdm1 | 31.4% | 31.6% | 32.0% | 30.7% | 28.5% | 26.2% | 24.9% | 18.8% |

**Fig. 1**. Base DER% for initial PLDA training experiments

train an ivector extractor on the AMI data, to use an ivector extractor taken from a separate speaker identification setup. The motivation for this was the relatively small amount of training data in the AMI corpus. As a result we used an ivector extractor that had been trained on a large amount of wideband data with many speakers. The training data was a combination of parts of the NIST 2008 SRE [13] training data, parts of Mixer 6 [14], and the VoxCeleb corpus [15].

#### *4.2.2. PLDA Training*

Due to the very limited number of speakers present in the AMI training data, we explored different approaches in handling the data to provide a better-performing PLDA. One approach shown to be successful in ASR systems is to 'reset' the ivectors periodically and treat the speech as coming from a new speaker [16]. In one set of experiments, we experimented with artificially multiplying the number of speakers by evenly splitting up ivectors for each speaker into a given number of 'sub-speakers' that were considered to be unique for the purpose of the training of the PLDA. We also took the approach of considering each speech segment to be a unique speaker, with each speaker label corresponding to the collection of short-term ivectors extracted within that segment.

We also conducted experiments in training a more powerful PLDA model. Due to the greater variance in ivectors for shorter time scale, we experimented with extracting long-term ivectors for the purpose of training the PLDA. The method we used was to pick a target duration as a floor for length, and then artificially concatenate speech segments from a speaker together until the new set of segments were all above the target duration in length.

We also reduced the number of ivectors used in the PLDA by passing only a fraction (referred to as the 'reduction factor') of a speaker's ivectors to the PLDA. We did this due to the small number of speaker present in the training data. We did not artificially increase the number of speakers as done with the short-term ivectors due to the long-term ivectors having less variance, which would confuse the PLDA with multiple ivectors from a single speaker being treated as separate speakers. As a result, we simply reduced the number of ivectors per speaker in hopes that the model would not overtrain on the small set of speakers and would thus generalize better. In addition, due to the decreased number of ivectors used in the training, we used a mean computed from the short-time ivectors for the global mean that was subtracted from the long-term ivectors.

The parameters for the long-term ivector target length and per-speaker reduction factor were tuned after the fact. Due to the already small test set, we used oracle calibration rather than a held-out set. However, the parameters were chosen using overall trends rather than the minimum error rate to minimize reporting statistical noise resulting from the size of the test set.

## 5. RESULTS

Figure 1 shows the results of our experiments with artificially increasing the number of speakers passed to the PLDA. We focused on the single far-field mic for these experiments. The baseline is equivalent to a speaker multiplier of 1. There are on average roughly 200 segments per speaker, so the segment-based training is given a number of speakers on par with the speaker multiplier of 200. Increasing the number of different speaker classes passed to the PLDA does decrease the error rate, with segment-based speaker labels giving the best error rate. As a result, we used this method of PLDA training for the other experiments on the differing speaker representations.

It is worth noting, however, that despite being passed identical ivectors with comparable number of sub-speaker classes, the segment-based training performs noticeably better than the case with a speaker multiplier of 200. We believe this to be the result of all the ivectors of a sub-speaker class being temporally close in the segment-based case in contrast to the speaker-multiplier case, where ivectors of a sub-speaker class are taken evenly throughout the recording. As a result, changes in a speaker's speech over the course of a recording may be simulating differences between speakers, leading to a better-trained PLDA.

| condition | baseline | external extractor |
|---|---|---|
| sdm1 | 18.8% | 14.3% |
| mdm8 | 15.1% | 9.9% |
| hms | 8.7% | 8.4% |

**Fig. 2**. DER% comparison between different speaker representations

Figure 2 shows the results of the experiments with different speaker representations. Using an ivector extractor trained on a greater amount of external data provides noticeable improvements to the error rate, with the exception of the synthetic near-field head mounted mic case. The modest im-

| | original extractor | | | | external extractor | | | |
|---|---|---|---|---|---|---|---|---|
| | baseline DER% | target ivec. length | reduction factor | experiment DER% | baseline DER% | target ivec. length | reduction factor | experiment DER% |
| sdm1 | 18.8% | 180s | 0.4 | 14.5% | 14.3% | 120s | 0.8 | 12.8% |
| mdm8 | 15.1% | 120s | 0.8 | 9.7% | 9.9% | 60s | 0.8 | 7.6% |
| hms | 8.7% | 120s | 0.8 | 8.0% | 8.4% | 60s | 0.8 | 4.8% |

**Fig. 3**. DER% comparison between baseline and long-term ivector PLDA training.

provements in this case suggest that the noise of a condition dictates the required quality of the model. The increase in performance using the external ivector extractor system suggests that providing an adequate amount of training data is the most critical component of performance of these systems.

In all cases, the far-field speech conditions performed worse than the synthetic near-field case, though the beamformed version approached the near-field case with the better systems. Since the conditions feature the exact same speech, this demonstrates that there is a degradation of diarization performance from near-field to far-field case, and steps taken to address this (e.g. beamforming) can improve performance.

Figure 3 shows the results of training the PLDA using the long-term ivectors, using both the baseline ivector system and the ivector system trained on external data. This method does improve the diarization error rate, though the gains on the better-trained are smaller. This does suggest, however, that better PLDA training could offset the harm done by a less well-trained ivector extractor, which could be useful in cases where there is a limited amount of in-domain training data.

In terms of the parameters used for the passing long-term ivectors to the PLDA model, the optimal operating region shifted towards longer segment lengths and more data reduction as the quality of the ivectors decreased, using the assumption that noisier data and worse extractor models lead to lower-quality ivectors. This makes intuitive sense, as using more frames per ivector would be expected to produce higher-quality ivectors which could offset a poor model. The reduction factor also could be mitigating the increased variance per speaker from poor-quality ivectors.

## 6. CONCLUSION

In this work we sought to establish the current state of diarization performance on far-field speech. For this we used the AMI Meeting Corpus, which introduced the additional problem of a small amount of training data. We have demonstrated that there is a degradation of diarization performance on far-field speech. Taking directed approaches to reducing this degradation shows gains, but standard methods have not yet reached state-of-the-art performance on near-field speech.

## 7. ACKNOWLEDGEMENTS

The bulk of research reported here was conducted at the 2017 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at Carnegie Mellon University and sponsored by Johns Hopkins University with unrestricted gifts from Amazon, Apple, Facebook, Google, and Microsoft, and we would like to thank them for making this work possible.

## 8. REFERENCES

[1] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4930–4934.

[2] W. Zhu and J. Pelecanos, "Online speaker diarization using adapted i-vector transforms," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5045–5049.

[3] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

[4] Grgor Dupuy, Sylvain Meignier, Paul Delglise, and Yannick Estve, "Recent improvements on ILP-based clustering for broadcast news speaker diarization," June 2014.

[5] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, Pierre Wellner, Steve Renals, and Samy Bengio, *The AMI Meeting Corpus: A Pre-announcement*, pp. 28–39, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[6] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, Sept 2007.

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

[8] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[9] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[10] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 413–417.

[11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Rue Marconi 19, Martigny, Dec. 2011, number Idiap-RR-04-2012, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[12] N. Brümmer and D. Garcia-Romero, "Generative modelling for unsupervised score calibration," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1680–1684.

[13] NIST Multimodal Information Group, "2008 NIST Speaker Recognition Evaluation Test Set LDC2011S08," Web download, Philadelphia: Linguistic Data Consortium, 2011.

[14] Linda Brandschain, David Graff, and Kevin Walker, "Mixer 6 Speech LDC2013S03," Hard drive, Philadelphia: Linguistic Data Consortium, 2013.

[15] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[16] Vijayaditya Peddinti, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," in *INTERSPEECH*, 2015.