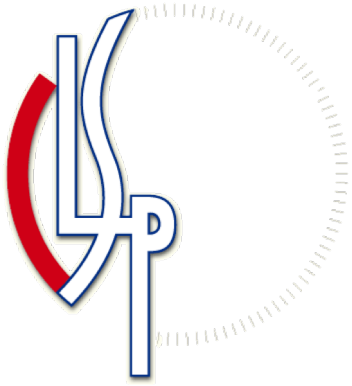
Lessons Learned from the **DIHARD Speech Diarization Challenge**

Sanjeev Khudanpur

Center for Language & Speech Processing, and
Human Language Technology Center of Excellence
Johns Hopkins University
July 19, 2018



human language technology
center of excellence

An upcoming INTERSPEECH paper

Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, Sanjeev Khudanpur

Center for Language and Speech Processing & Human Language Technology Center of Excellence
Johns Hopkins University, USA

gsell@jhu.edu

Abstract

We describe in this paper the experiences of the Johns Hopkins University team during the inaugural DIHARD diarization evaluation. This new task provided microphone recordings in a variety of difficult conditions and challenged researchers to fully consider all speaker activity, without the currently typical practices of unscored collars or ignored overlapping speaker segments. This paper explores several key aspects of currently state-of-the-art diarization methods, such as training data se-

ically estimate marks with a speech activity detection (SAD) algorithm.

This paper describes the submissions for the inaugural DIHARD challenge from the Johns Hopkins University (JHU) team, as well as our experiments on the path from an initial system built for Callhome diarization to our final microphone diarization system. The discussion also includes possible directions for future work, as the limited time of the challenge meant many paths were necessarily left unexplored.

The First DIHARD Speech Diarization Challenge

DIHARD is a new annual challenge focusing on "hard" diarization; that is, speech diarization for challenging corpora where there is an expectation that the current state-of-the-art will fare poorly, including, but not limited to:

- clinical interviews
- extended child language acquisition recordings
- YouTube videos
- "speech in the wild" (e.g., recordings in restaurants)

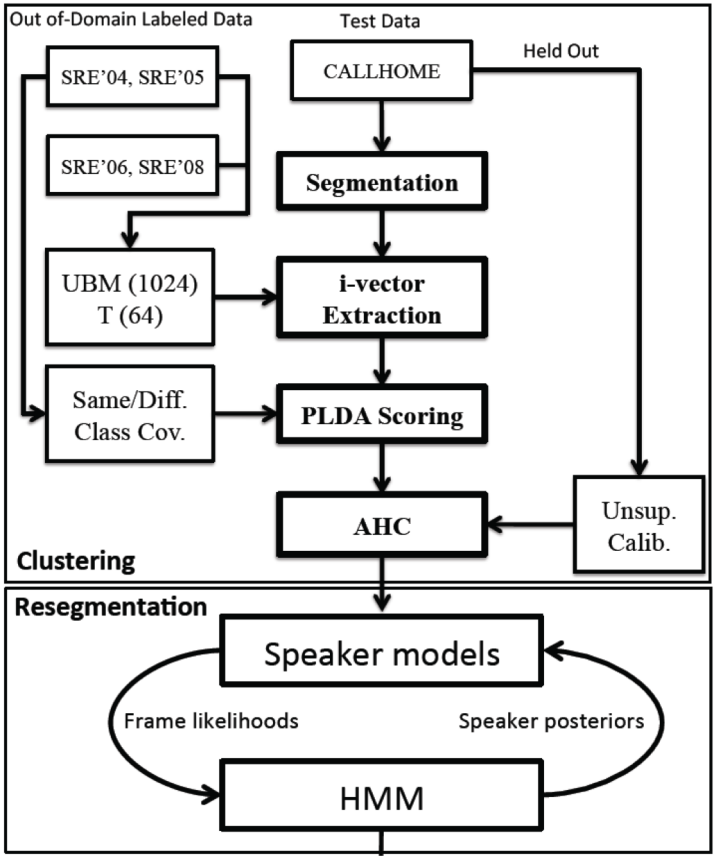
Because the performance of a diarization system is highly dependent on the quality of the speech activity detection (SAD) system used, the challenge will have two tracks:

- Track 1: diarization beginning from gold speech segmentation
- Track 2: diarization from scratch

The results of this initial challenge will be presented at a special session at Interspeech 2018 in Hyderabad.

Our “out of the box” solution

- Perform speech activity detection
  - CURRENNT trained on Switchboard
- Extract i-vectors for 2 sec segments
  - Fixed width, overlapped: stride 1 sec
- Estimate/Apply PLDA
  - Training: Estimate PLDA parameters
  - Test: Compute pairwise similarity
- Cluster 2 sec segments by speaker(s)
- Calibrate stopping threshold
  - Supervised: on held out recordings
  - Unsupervised: on test recordings
- Optionally, refine segment boundaries



SPEAKER DIARIZATION WITH PLDA I-VECTOR SCORING AND UNSUPERVISED CALIBRATION

Gregory Sell and Daniel Garcia-Romero

Our “out of the box” solution

Diarization Method	Track 1 DER		Track 2 DER	
	Dev Set	Eval Set	Dev Set	Eval Set
Declare there’s only 1 speaker!	36.0%	39.0%	48.7%	55.9%
“Out of the Box” (CALLHOME)	26.7%	31.6%	40.9%	50.8%

Improving speech activity detection

- CURRENNT SAD
  - Recurrent neural network
    - https://sourceforge.net/projects/currennt/
  - Trained on Switchboard
    - Plus data augmentation
  - 8 kHz speech, 13 MFCCs
  - 10.2% Miss, 4.6% FA
    - Higher FA worse for DER
- 5-Layer TDNN SAD
  - Feed forward, 640 ms span
  - Trained on Europarl
    - Plus data augmentation
  - 16 kHz speech, 24 MFCCs
  - 17.4% Miss, 4.8% FA
- Fine-tune on DIHARD
  - 7.3% Miss, 4.1% FA
  - Fine-tune on each domain
    - 6.1% Miss, 4.2% FA

Wideband data for i-vectors & PLDA

Diarization Method	Track 1 DER		Track 2 DER	
	Dev Set	Eval Set	Dev Set	Eval Set
Declare there's only 1 speaker!	36.0%	39.0%	48.7%	55.9%
"Out of the Box" (CALLHOME)	26.7%	31.6%	40.9%	50.8%
i-vectors, 16 kHz data, no VB	21.7%	28.1%	33.7%	40.4%

Using x-vectors instead of i-vectors

Diarization Method	Track 1 DER		Track 2 DER	
	Dev Set	Eval Set	Dev Set	Eval Set
Declare there's only 1 speaker!	36.0%	39.0%	48.7%	55.9%
"Out of the Box" (CALLHOME)	26.7%	31.6%	40.9%	50.8%
i-vectors, 16 kHz data, no VB	21.7%	28.1%	33.7%	40.4%
x-vectors, 16 kHz data, no VB	20.0%	25.9%	31.8%	39.4%

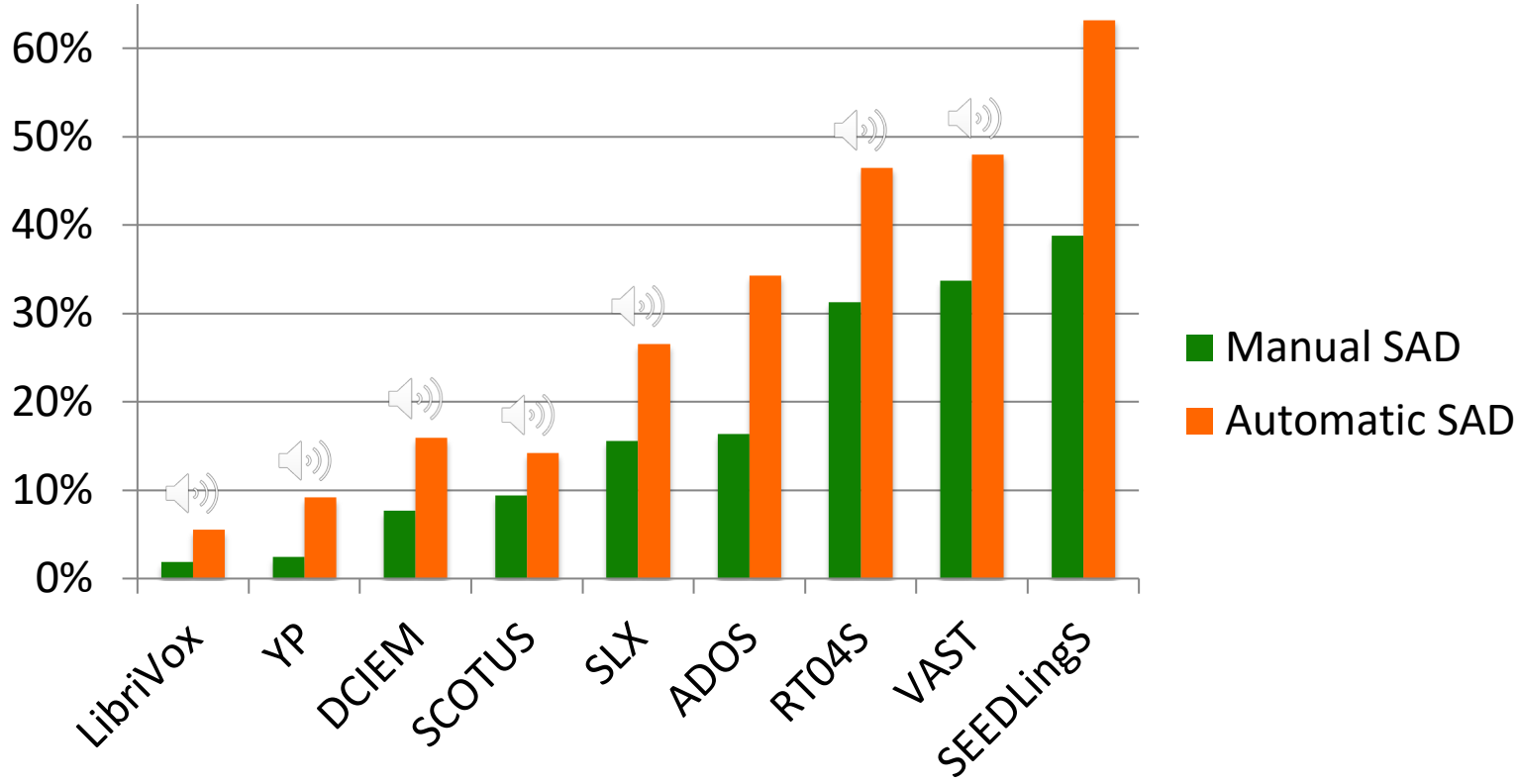
Improving resegmentation

Diarization Method	Track 1 DER		Track 2 DER	
	Dev Set	Eval Set	Dev Set	Eval Set
Declare there's only 1 speaker!	36.0%	39.0%	48.7%	55.9%
"Out of the Box" (CALLHOME)	26.7%	31.6%	40.9%	50.8%
i-vectors, 16 kHz data, no VB	21.7%	28.1%	33.7%	40.4%
x-vectors, 16 kHz data, no VB	20.0%	25.9%	31.8%	39.4%
i-vectors, 16 kHz data, with VB*	19.7%	25.1%	31.3%	37.4%
x-vectors, 16 kHz data, with VB*	18.2%	23.7%	29.8%	37.3%

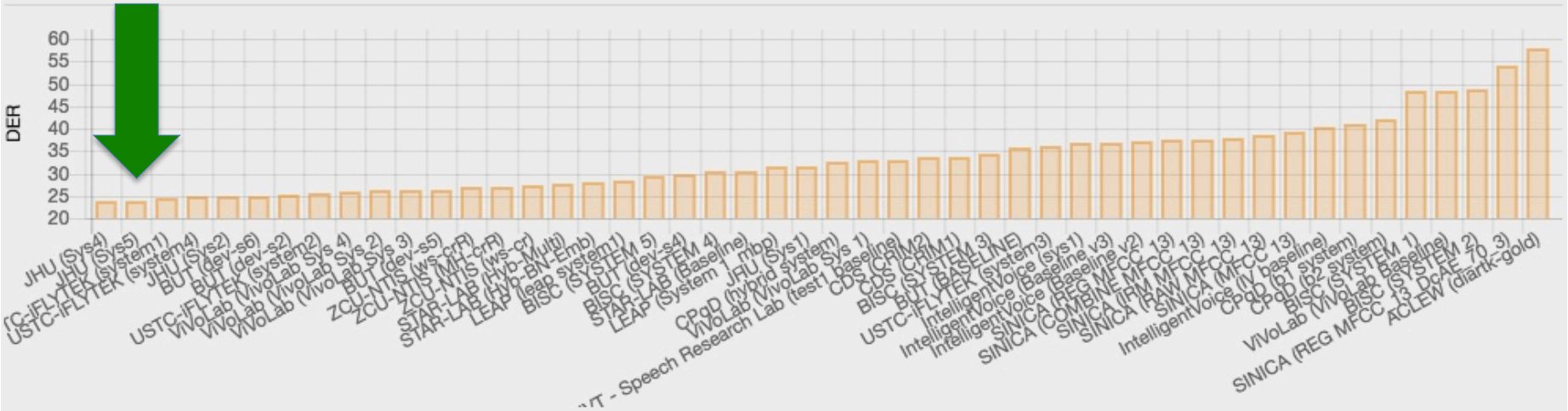
Fusing PLDA scores from (i|x)-vectors

Diarization Method	Track 1 DER		Track 2 DER	
	Dev Set	Eval Set	Dev Set	Eval Set
Declare there's only 1 speaker!	36.0%	39.0%	48.7%	55.9%
"Out of the Box" (CALLHOME)	26.7%	31.6%	40.9%	50.8%
i-vectors, 16 kHz data, no VB	21.7%	28.1%	33.7%	40.4%
x-vectors, 16 kHz data, no VB	20.0%	25.9%	31.8%	39.4%
i-vectors, 16 kHz data, with VB*	19.7%	25.1%	31.3%	37.4%
x-vectors, 16 kHz data, with VB*	18.2%	23.7%	29.8%	37.3%
(i+x)-vector fusion, 16 kHz, VB*	18.2%	24.0%	30.3%	37.2%

Corpus-wise breakdown of DER



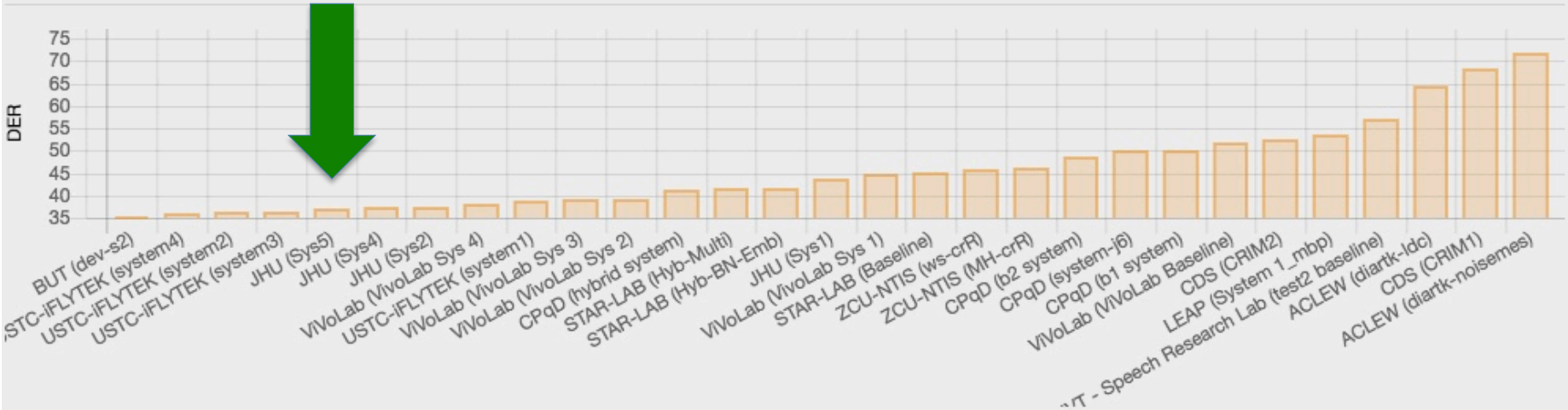
Track1



Bar | Scatter | Off

Rank	Team (System)	Team Name	System Name	Date	DER	MI
1	JHU (Sys4)	JHU	Sys4	2018-03-22 18:47:24	23.73 %	8.4400
2	JHU (Sys5)	JHU	Sys5	2018-03-22 18:48:42	23.99 %	8.4300
3	USTC-iFLYTEK (system1)	USTC-iFLYTEK	system1	2018-03-23 19:00:48	24.56 %	8.4700
4	USTC-iFLYTEK (system4)	USTC-iFLYTEK	system4	2018-03-22 16:06:59	24.96 %	8.4600
5	JHU (Sys2)	JHU	Sys2	2018-03-22 18:18:36	25.06 %	8.4200

Track2



Bar | Scatter | Off

Rank	Team (System)	Team_Name	System_Name	Date	DER	MI
1	BUT (dev-s2)	BUT	dev-s2	2018-03-23 18:21:08	35.51 %	8.0700
2	USTC-iFLYTEK (system4)	USTC-iFLYTEK	system4	2018-03-23 22:15:12	36.05 %	8.0800
3	USTC-iFLYTEK (system2)	USTC-iFLYTEK	system2	2018-03-20 16:49:21	36.39 %	8.0900
4	USTC-iFLYTEK (system3)	USTC-iFLYTEK	system3	2018-03-23 22:56:23	36.56 %	8.0600
5	JHU (Sys5)	JHU	Sys5	2018-03-22 18:48:42	37.19 %	8.0400

Take-home message(s) about diarization

- A good “out of the box” CALLHOME system needed considerable tuning to work reasonably on DIHARD
  - Using bandwidth-matched data was very helpful for training the (i|x)-vector extractor & PLDA parameters
  - Corpus-specific tuning was helpful for some modules
  - Efforts at speech enhancement haven’t helped … so far
- SAD still requires some serious improvement
- VB re-segmentation is somewhat fragile
- Several corpora are still terribly HARD for diarization