Speech Recognition Israel

Meetups

Technical Facebook group!

Community

Diarization in practice

Yoav Ramon

ML Engineer at Hi.Auto

Yoav@hi.auto

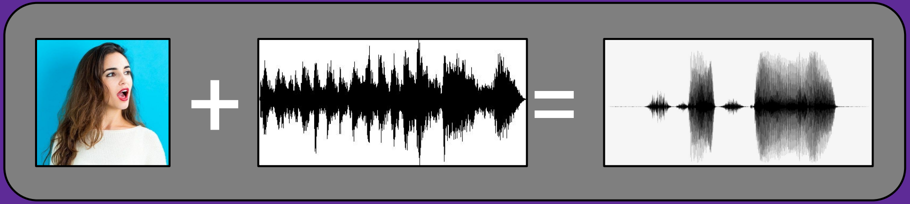
in Yoav Ramon

@YoavR7

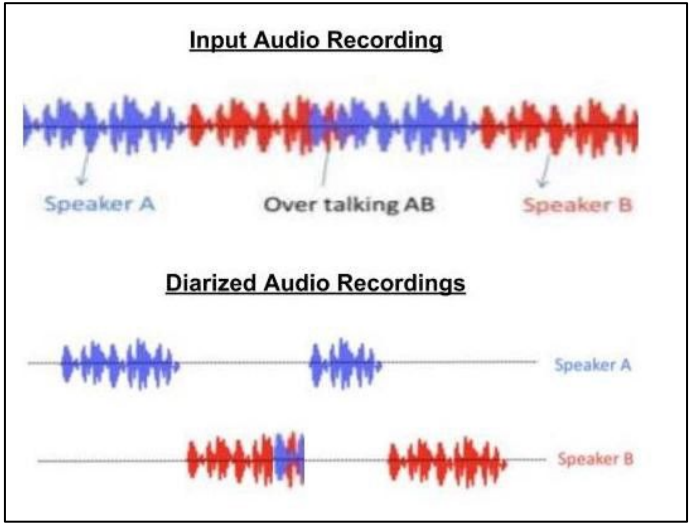
M @yoavramon

Hi Auto

Audio-Visual Robust Speech Solution



The Problem



Why it's important?

- Speaker identification
- Speech Recognition
- Real-world data analysis

"Who spoke when"

Diarization ≠ Speaker Separation

How good are we?

CASE SOLVED

Diarization with X-vectors
(Snyder, 2018)

	DER
Without Oracle	8.39
With Oracle	7.12

Callhome Challenge

1997, 120 Calls, 2 Channels, Telephony
60 Hours +-

So... Is it solved?

> “While state-of-the-art diarization systems perform remarkably well for some domains (e.g., conversational telephone speech such as CallHome), **as was discovered at the 2017 JSALT Summer Workshop at CMU, this success does not transfer to more challenging corpora** such as child language recordings, clinical interviews, speech in reverberant environments, web video, and **speech in the wild**” *(Church et al., Feb. 2018)*

DIHARD Challenge

Interspeech 2018, September

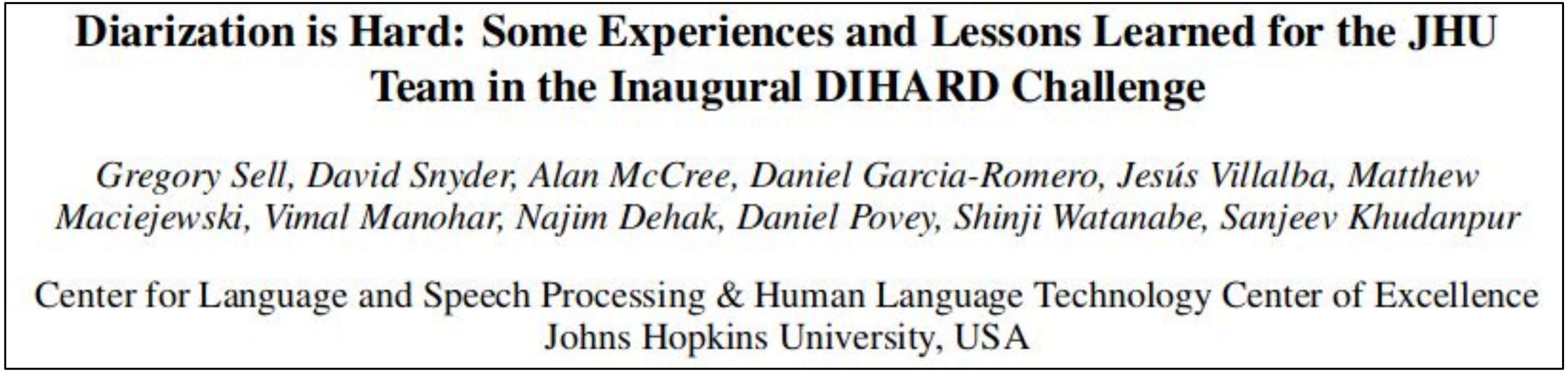
Development Test

- Only 19 Hours
- 9 Domains
  (*Child language, Supreme Court, Clinical interviews, Radio interviews, Map tasks, Sociolinguistic interviews, Meeting speech, Audiobooks, YouTube videos)*
- Single Channel
- 5 Minutes per sample

Evaluation Test

- 21 Hours
- 3 Different domains
  (*Sociolinguistic interviews, Meeting speech, Restaurant conversation)*
- Single Channel
- 5 Minutes per sample

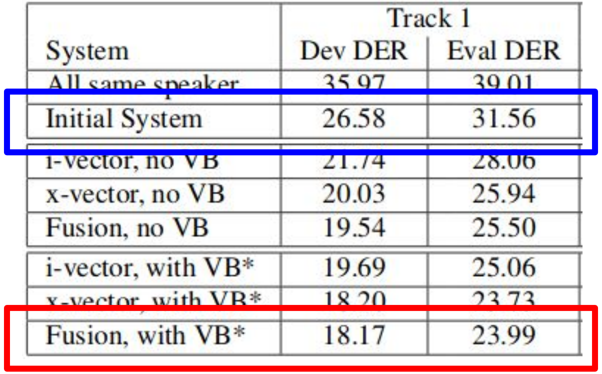
So… Is it solved?



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



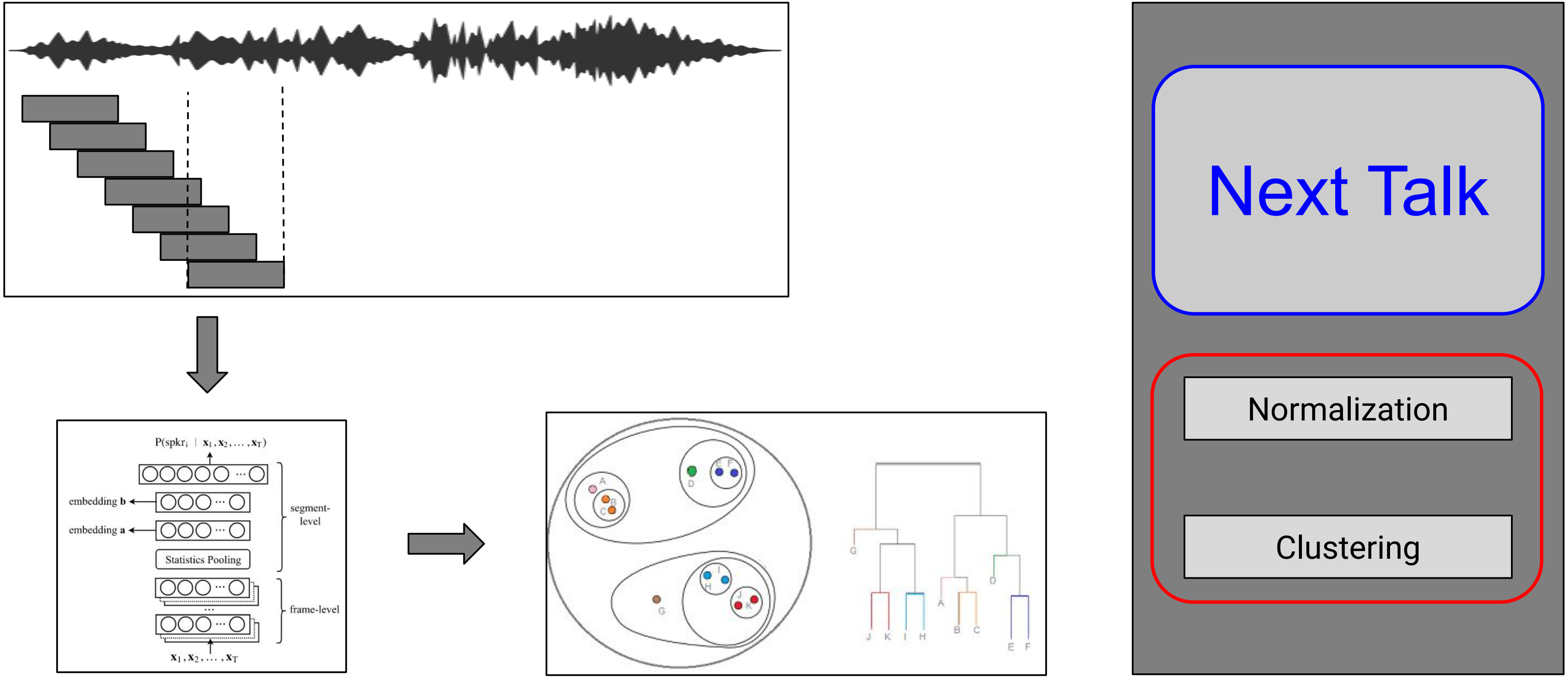
System	Track 1 Dev DER	 Eval DER
All same speaker	35.97	39.01
Initial System	26.58	31.56
i-vector, no VB	21.74	28.06
x-vector, no VB	20.03	25.94
Fusion, no VB	19.54	25.50
i-vector, with VB*	19.69	25.06
x-vector, with VB*	18.20	23.73
Fusion, with VB*	18.17	23.99

Far from being solved...

Implementation

(So…. How do I do it?)

The framework:

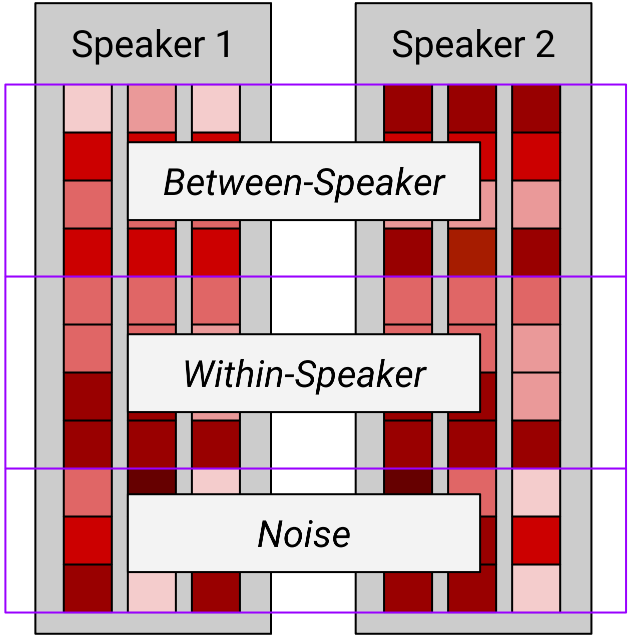


Next Talk

Normalization

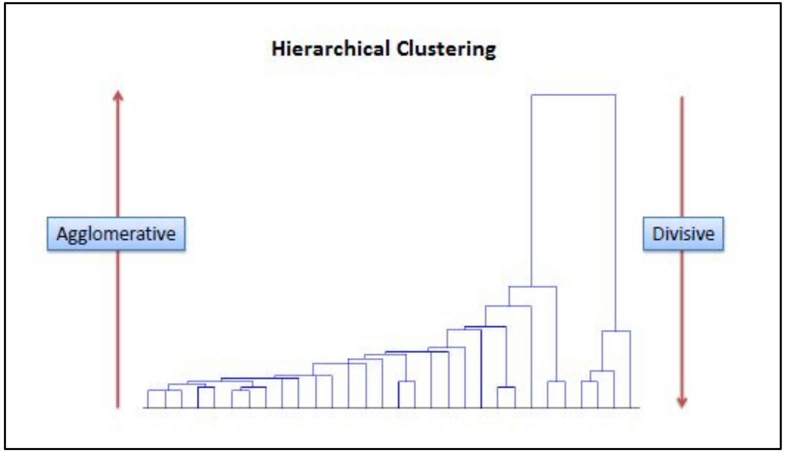
Clustering

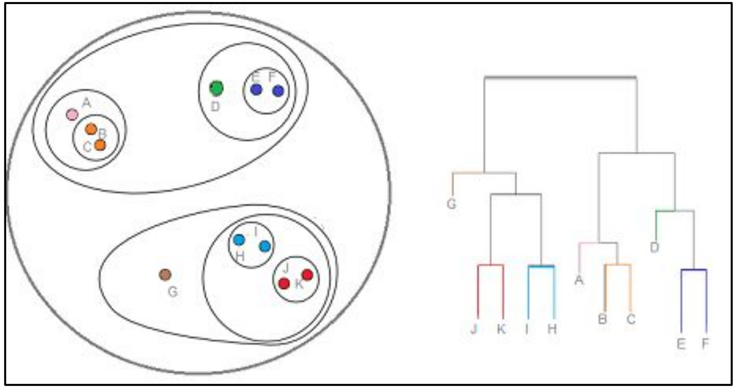
PLDA Normalization

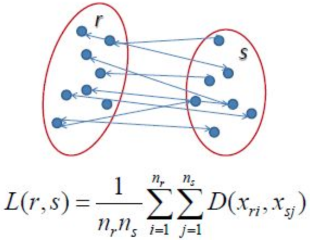


$$\Phi = \mu + Vy + Ux + \epsilon$$

Agglomerative Clustering





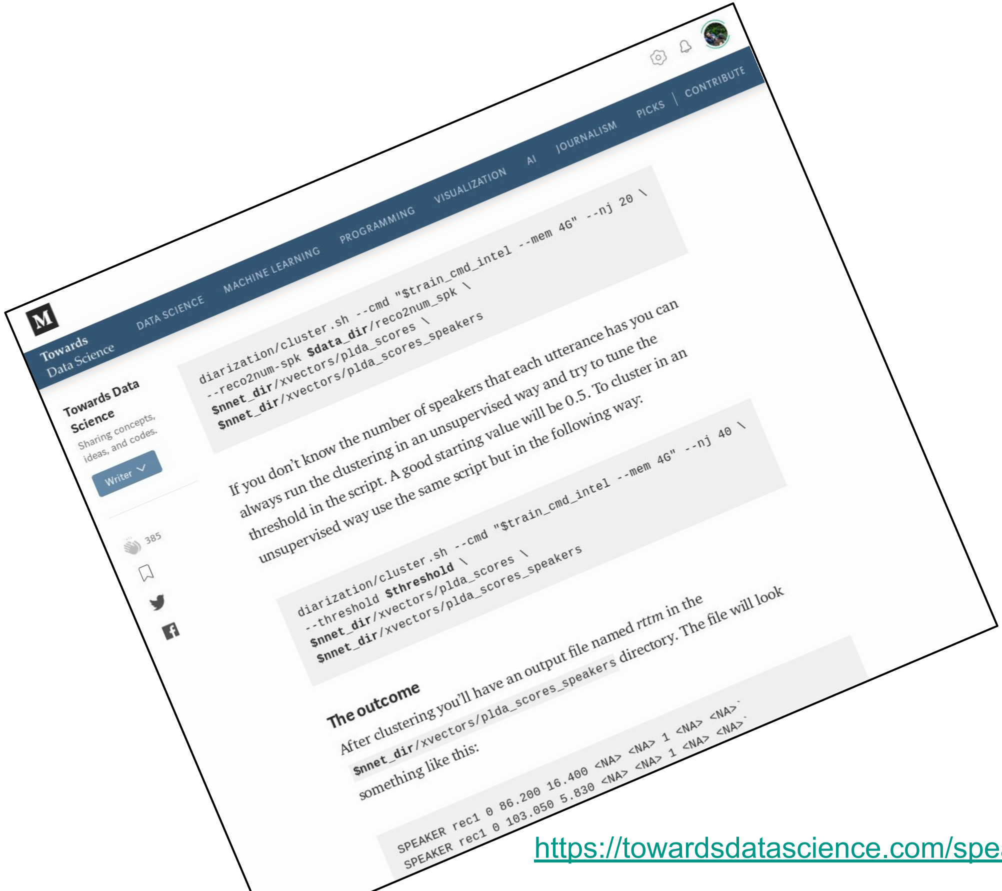


$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

With Oracle: Until we have enough speakers

Without Oracle: With Threshold

You can do it with Kaldi!



```
diarization/cluster.sh --cmd "$train_cmd_intel --mem 4G" --nj 20 \
  --reco2num-spk $data_dir/reco2num_spk \
  $nnet_dir/xvectors/plda_scores \
  $nnet_dir/xvectors/plda_scores_speakers
```

If you don't know the number of speakers that each utterance has you can always run the clustering in an unsupervised way and try to tune the threshold in the script. A good starting value will be 0.5. To cluster in an unsupervised way use the same script but in the following way:

```
diarization/cluster.sh --cmd "$train_cmd_intel --mem 4G" --nj 40 \
  --threshold $threshold \
  $nnet_dir/xvectors/plda_scores \
  $nnet_dir/xvectors/plda_scores_speakers
```

The outcome

After clustering you'll have an output file named *rttm* in the $nnet_dir/xvectors/plda_scores_speakers directory. The file will look something like this:

```
SPEAKER rec1 0 86.200 16.400 <NA> <NA> 1 <NA> <NA>
SPEAKER rec1 0 103.050 5.830 <NA> <NA> 1 <NA> <NA>
```

- It's easy to use
- There are pretrained models
  - Although you'll want to adapt them
- There is support in clustering with oracle.
- It's easy to integrate inside speech processing pipeline

https://towardsdatascience.com/speaker-diarization-with-kaldi-e30301b05cc8

Challenges

- Real Time
- Not good enough vectorization
- Cross-Domain robustness (20% DER)
- Estimating the number of speakers might be hard
- Dimensionality reduction might create too-sparse vectors

Advantages

- Really Active research…
- Not so hard to implement.
- “Building Block” architecture leaves place to innovation.
- The value to ASR Systems is huge, even with relatively poor DER
- “Transfer Learning” on PLDA is easy

Hi Auto

Thank you!

Questions?