Monty Python Speech Diarization

Patricio Foncea, Andreea Georgescu, Andrey Sushko

12/10/2018

Problem Formulation

Data Available

- Flying Circus sketches, audio and subtitles.
- Each data point represents a line in a sketch
  Specifically, the averaged spectrogram data for the time interval of the line.
- Small sample of lines with labeled speaker.

Goal

- Assign lines to speakers.
- No prior information about different speakers.
- A version of the second stage in speech diarization:
  Once change in speaker candidates are identified, determine whether speech segments belong to same speakers.

Motivation

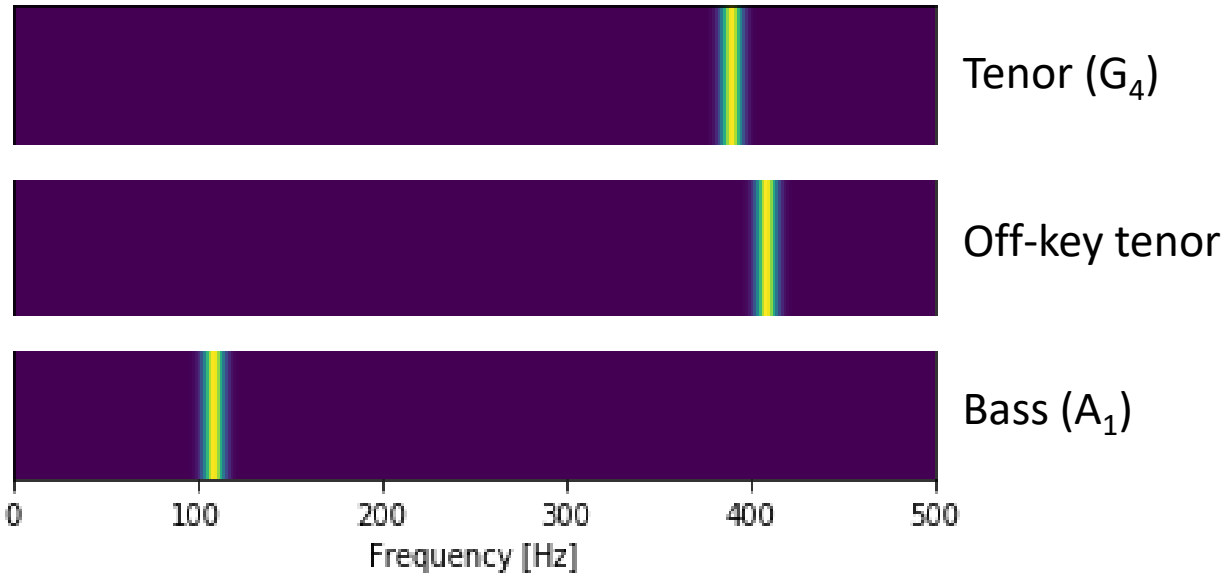
- Original goal: use subtitle files to generate new sketches (Neural Nets).
  Hypothesis: given time delimitation of lines, use clustering to supplement speaker to the data.
- Other application: determine speaker without storing content.

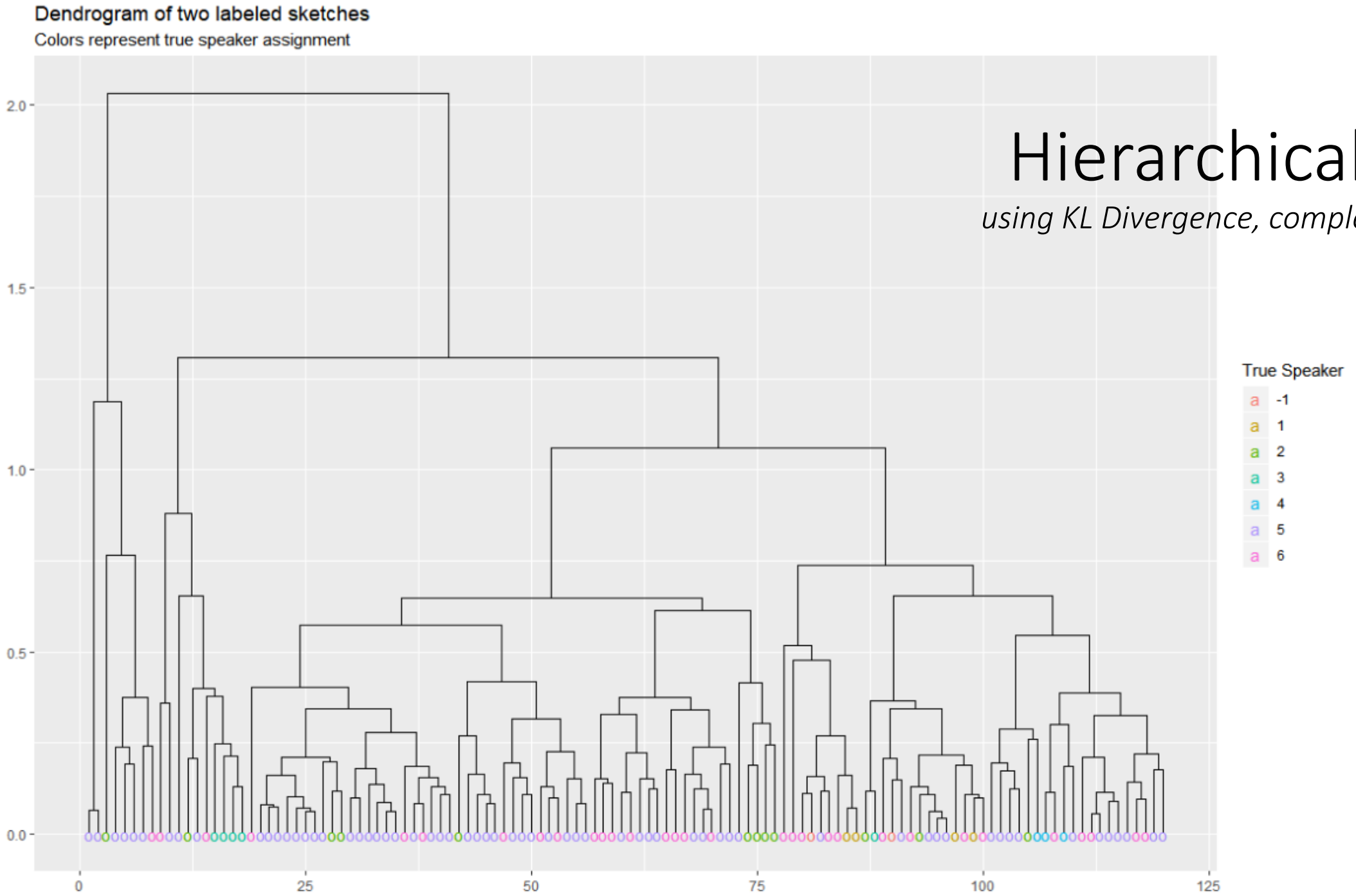
Methods – overview

- Unsupervised Clustering
  - Naïve k-means
  - Hierarchical trees with custom distances
    - Earth moving distance and KL Divergence
- Supervised Approach
  - For two lines, predict whether they belong to the same speaker
  - Fit logistic regression on labeled sample
- PCA and Gaussian Mixture
  - Run PCA to reduce dimensionality of frequency bins to a few significant chords
  - Estimate a gaussian mixture model to cluster lines

Naïve k-means

- Bad performance since clusters are not spherical.
  - Coordinates / features have a meaningful order
  - In this sense, we can think of data points more like distributions
  - Look at appropriate distances:
    - Averaged KL divergence
    - Earth moving distance



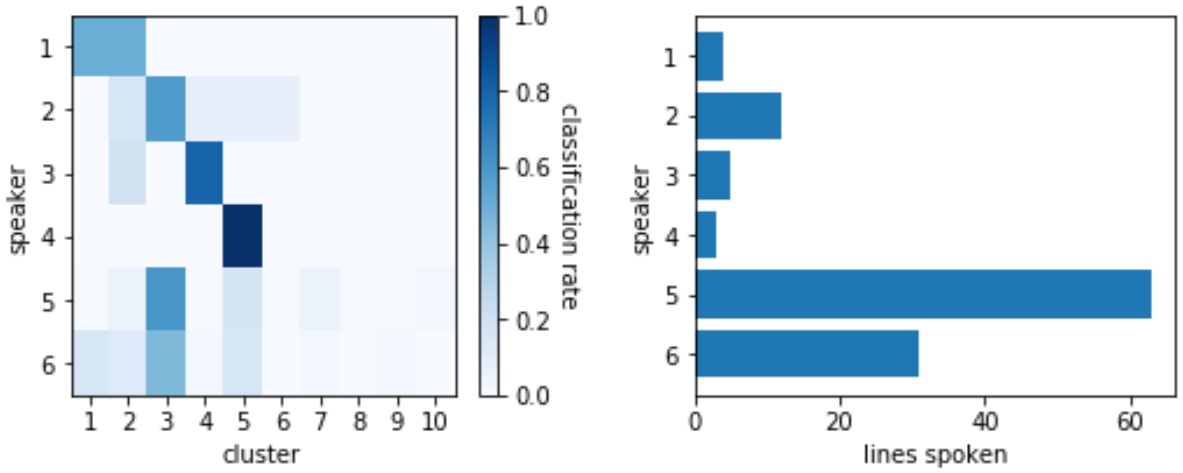


Hierarchical trees

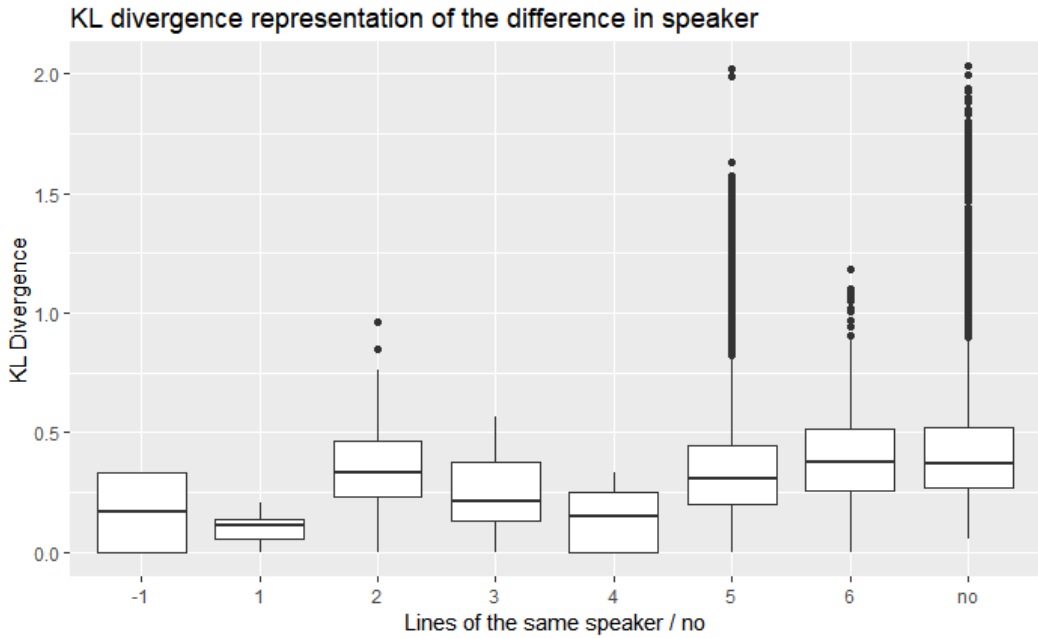
using KL Divergence, complete clustering

Hierarchical Trees

Accuracy of clustering when we cut the tree before at 10 clusters

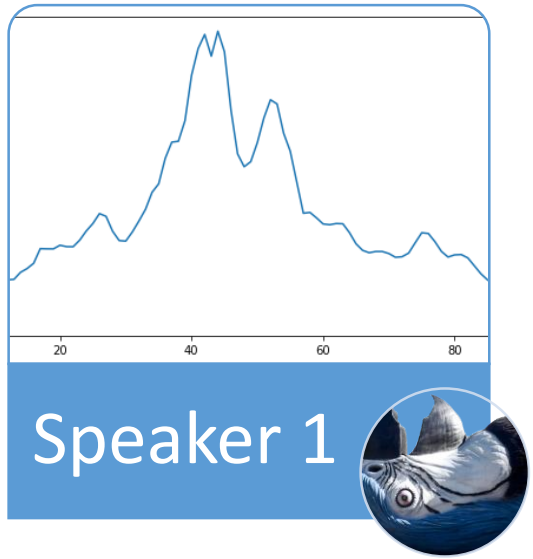


Pairwise distance distribution across pairs in the same / different clusters

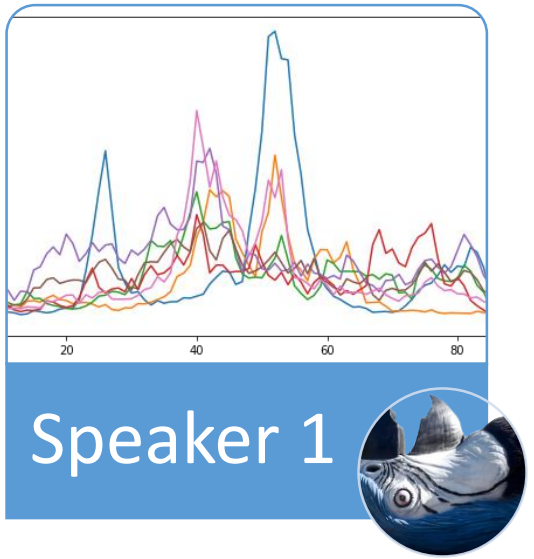


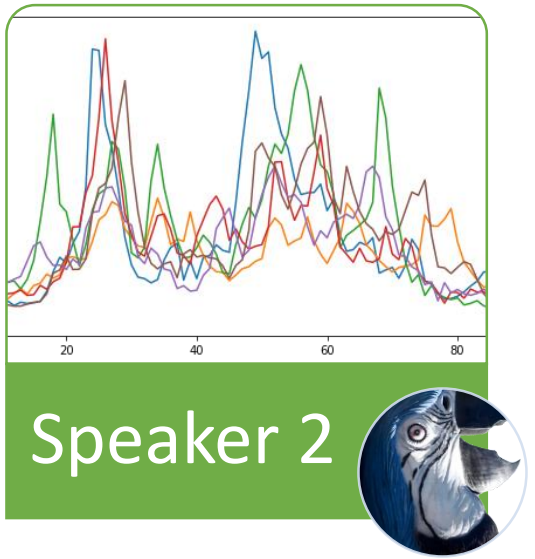
Why do our distances perform badly?

Looking on average data per speaker, distances chosen look promising.



Each data point from one speaker is a poor estimation of complex distribution





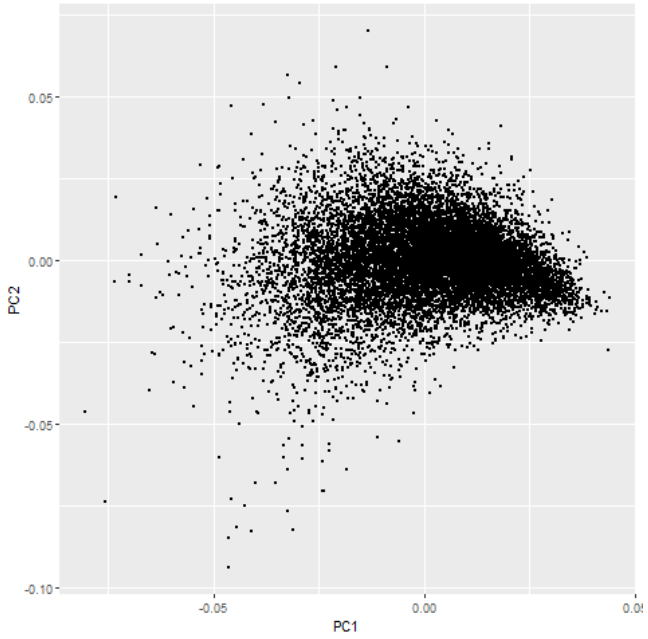
Supervised Approach

- Approach: for a pair of data points, learn whether they belong to the same speaker or not.
  - Used logistic regression on labeled sample.
  - When Train = Test: 82% accuracy
  - When Train ≠ Test: 30% accuracy

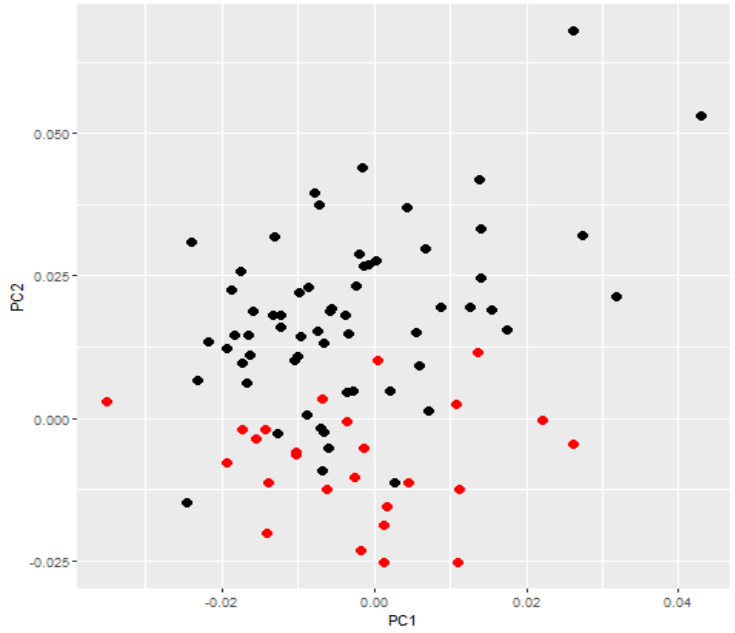
- Conclusion:
  - Labeled data is small and not representative of entire dataset
  - With more representative labeled data, this method could work well
    - Especially since it looks on entire sample from one speaker.

Feature Selection and Dimensionality Reduction

- Ran PCA on entire data set
- Data is separable but not cluster-able
- Suggests supervised method would work well with enough data



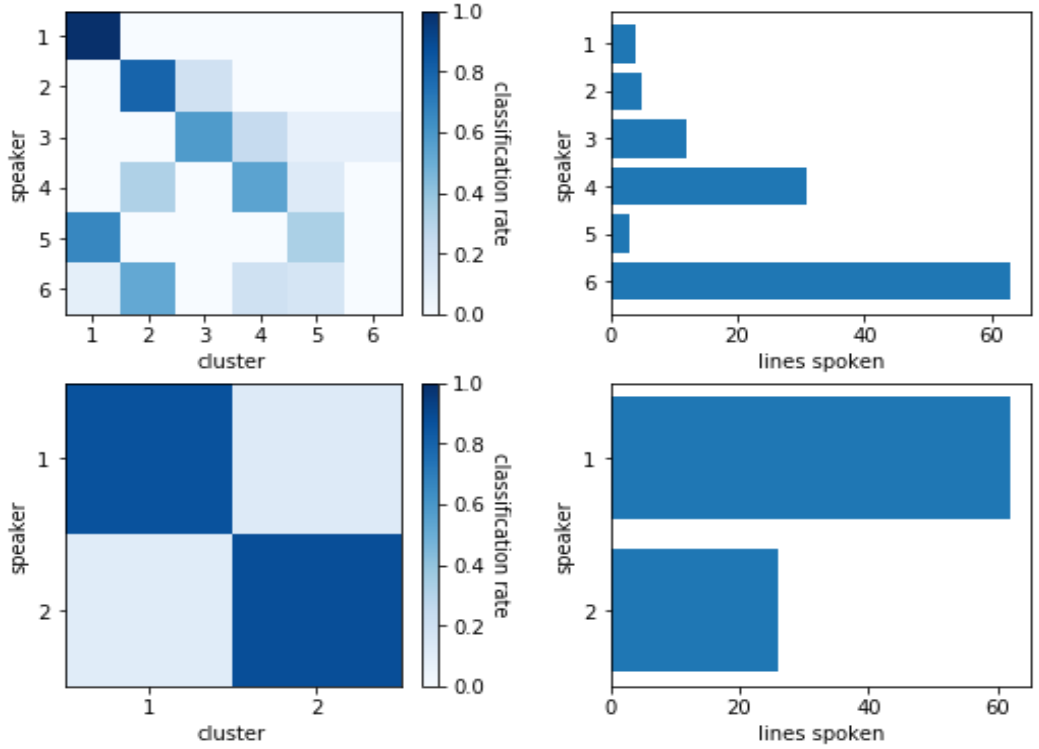
Entire dataset in first two PCs



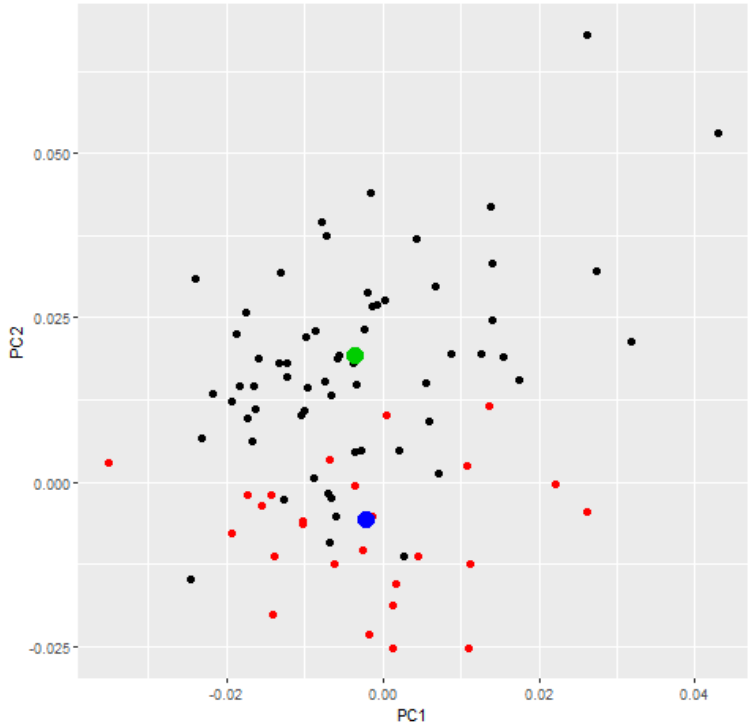
Labeled data with two speakers in first two PCs

Gaussian Mixture Model

- Fit model on labeled sample



Accuracy of Mixture of Gaussian



Projection onto components

And Now for Something Completely Different

Script generation via char-rnn

174
00:10:40,207 --> 00:10:42,740
I don't know what I mean,
lord the strike.

223
00:10:27,591 --> 00:10:30,083
Well, we were the country

256
00:10:56,990 --> 00:10:59,163
I didn't want to make
the road of the wood.

273
00:14:12,286 --> 00:14:14,868
I can see the late
in the head of the world to me.

223
00:14:22,863 --> 00:14:25,923
I think you go to the
present the name of the collect.

240
00:14:02,646 --> 00:14:04,678
I was a special struck
of the look in the studio

253
00:14:28,161 --> 00:14:30,244
Yes, yes, it's a little bit,
by a bloody little back.

251
00:10:38,221 --> 00:10:41,512
The last week of the present scene

254
00:14:03,207 --> 00:14:04,905
I'm not a problems
and believe it again.

284
00:14:15,896 --> 00:14:17,698
The great address the stranges

296
00:14:37,811 --> 00:14:41,982
What a stupid.