



# Acoustic Scene Classification for Speaker Diarization

Supervised by Md Sahidullah & Prof. Romain Serizel

---

Tahani FENNIR - Fatima HABIB - Cécile MACAIRE

Tuesday, 2 June 2020

University of Lorraine, IDMC

# Table of contents

1. **The Aim of the Project**
2. **Speaker Diarization: an introduction**
  - 2.1 Definition & Applications
  - 2.2 Challenges & Problem Statement
  - 2.3 Results of the development dataset
  - 2.4 t-SNE visualization
3. **Acoustic Scene Classification for Speaker Diarization**
  - 3.1 Unsupervised method: Clustering
    - 3.1.1 Applying K-means on x-vectors data set
    - 3.1.2 Applying diarization on the new classification
    - 3.1.3 Applying K-means on i-vectors data set
    - 3.1.4 Applying diarization on the new classification
  - 3.2 Supervised method: Classification
4. **Conclusion**
5. **Future prospects**

# 1. The Aim of the Project

---

# 1. The Aim of the Project

Investigate speaker diarization performance in the presence of different acoustic environments (with DIHARD II challenge dataset [5]), which involve applying different classification methods.

## 2. Speaker Diarization: an introduction

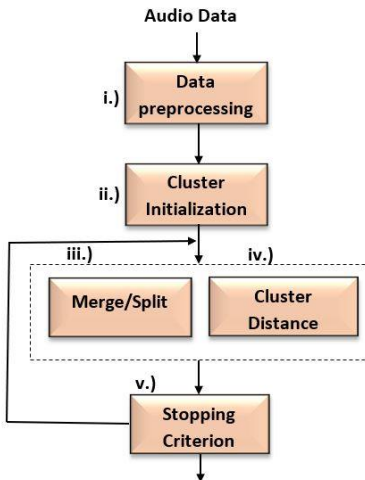
---

## 2.1 Definition & Applications

### Definition:

- Speaker diarization is the operation of labeling a speech signal with labels corresponding to the identity of speakers.
- It is the job of deciding "who spoke when?" in an audio or video recording involving an estimated number of speakers and an unspecified number of speakers.
- It has become a key technology for many tasks, including navigation, retrieval, or higher-level audio data inference.

## 2.1 Definition & Applications



**Figure 1:** General speaker diarization architecture

## 2.1 Definition & Applications

There are **three main application domains** for speaker diarization, as shown by Reynolds and Torres-Carrasquillo (2004):

- Broadcast news (BN): radio and TV programs with a variety of content
- Meetings: meetings or lectures in which multiple individuals communicate in the same room.
- Conversational telephone speech (CTS): single-channel recording of telephone conversations between two or more individuals.



## 2.2 Challenges & problem statement

**The biggest single challenge is** the handling of overlapping speech, which needs to be attributed to multiple speakers. Specifically, in regions in which more than one speaker is active, missed speech errors will be incurred and given the high performance of some state-of-the-art systems, this can be a substantial fraction of the overall diarization error.

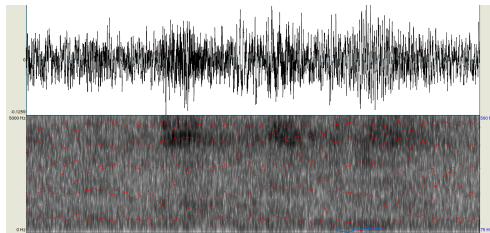
The fields of application, from broadcast news, to lectures and meetings, vary widely and pose different problems, such as

- Access to multiple microphones,
- Multimedia information,
- Recognize the location and acoustic sign or overlapping speech.

The detection and treatment of overlapping speech remains an unresolved problem.

## 2.2 Challenges & problem statement

The diarization can also be of poor quality when the speech is recording in a noisy environment. A poor diarization can conduct to misidentification of the speaker, or an absence of speech identification (Sounds can be identified as speech when they're not. We can take the example of child noises, such as children babbling.)



**Figure 2:** Spectrogram and Formants of a speech in a noisy environment (audio file taken from the DIHARD II challenge in the category restaurant)

## 2.2 Challenges & problem statement

In the period of the spring 2018, the initial DIHARD challenge ran.

**DIHARD I [Ryant et al. 2018]:** contains a one channel input condition using wide band speech sample from 11 demanding fields, ranging from clean recordings of read audio books to very noisy high interactive recording of speech.

**The DIHARD II challenge [Ryant et al. 2019]:** is the second in a series of speaker diarization challenges and was designed to enhance the robustness of diarization systems to variation in recording equipment, noise levels and conversational environments.

## 2.2 Challenges & problem statement

Showing in further details the problem statement of speaker diarization, for each 11 categories which constitute the DIHARD II challenge dataset.

The DIHARD II challenge dataset 11 categories are as follow:

- audiobooks,
- broadcast interviews,
- child,
- clinical,
- court,
- map tasks,
- meeting,
- restaurant,
- socio-field (everyday conversations),
- socio-lab,
- webvideos.

## 2.2 Challenges & problem statement

**DER** (Diarization Error Rate) is the overall percentage of the reference speaker time that is not accurately attributed to the speaker. The accurately attributed reference speaker time is described in terms of optimal mapping between the reference speakers and the system speakers. More specifically, the DER is equal to

$$DER = \frac{FA + MISS + ERROR}{TOTAL}$$

**JER** (Jaccard Error Rate) a metric based on Jaccard Index and specially computed for the DIHARD II challenge. The Jaccard Index is used to calculate the accuracy of a segmentation thanks to the ratio between two segmentations.

$$JER = \frac{FA + MISS}{TOTAL}$$

## 2.3 Results of the development dataset

Categories	Average DER	Average JER	Number of files
audiobooks	2.8	2.82	12
broadcast interviews	5.68	41.08	12
child	31.79	61.74	23
clinical	20.83	36.23	24
court	15.19	56.37	12
maptask	6.59	11.85	23
meeting	34.56	61.09	14
restaurant	49.77	79.03	12
socio field	14.84	40.19	12
socio lab	10.94	16.32	16
webvideo	37.85	62.15	32
<b>TOTAL</b>	<b>23.16</b>	<b>55.75</b>	<b>192</b>

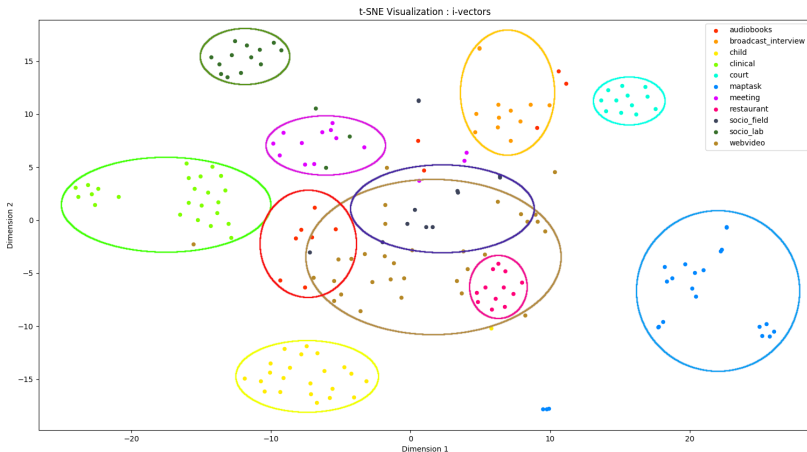
**Table 1:** The development dataset consists of 192 files. The files are splitted into different categories.

## 2.3 Results of the development dataset

Categories	DER	JER	Threshold
audiobooks	0.00	0.00	- 2.0
broadcast interviews	7.07	47.47	- 0.8
child	36.10	67.05	- 0.4
clinical	17.80	28.76	- 0.3
court	13.23	47,54	0.0
maptask	8.58	15.20	- 0.4
meeting	33.11	60.53	- 0.4
restaurant	52.00	77.82	- 0.3
socio field	24.48	55.66	- 0.5
socio lab	11.26	18.75	- 0.6
webvideo	39.57	77.74	- 0.5
<b>TOTAL</b>	<b>22.10</b>	<b>45.13</b>	-

**Table 2:** Table with the DER and JER scores for each category with the best threshold computed by Kaldi tool for each. The number of files is the same.

## 2.4 t-SNE visualization



**Figure 3:** t-SNE visualization of audio recordings of DIHARD II dataset with i-vectors. We can see that some files are clearly not well clustered.



### 3. Acoustic Scene Classification for Speaker Diarization

---

## 3.1 Unsupervised method: Clustering

*Unsupervised machine learning algorithms derive patterns from a dataset without reference to known, or labeled, outcomes.*

K-Means:

- Randomly pick K centroids that are used as starting points for each cluster.
- Conducts iterative (repetitive) calculations to optimize the position of the centroids.
- It stops creating and optimizing clusters when either:
  - The centroids have stabilized — their values have not changed because the clustering has been efficient. OR
  - The number of iterations defined has been accomplished.

## 3.1 Unsupervised method: K-means

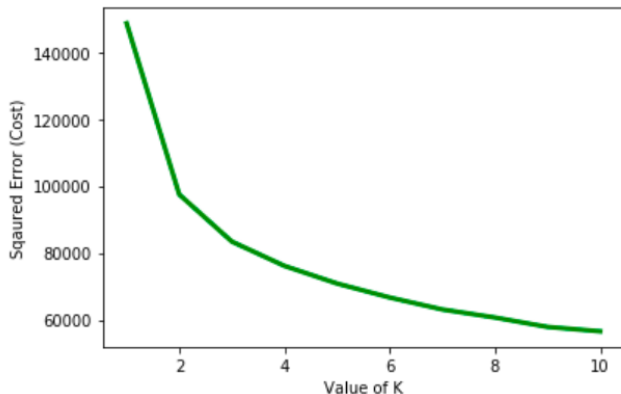
Application of the k-means algorithm on the DIHARD II challenge dataset

### **Constructing the K-means model**

1. Select the best K, using elbow method.
2. Cluster the data using the k-means model.
3. Evaluate the model.

### 3.1.1 Applying K-means on x-vectors data set

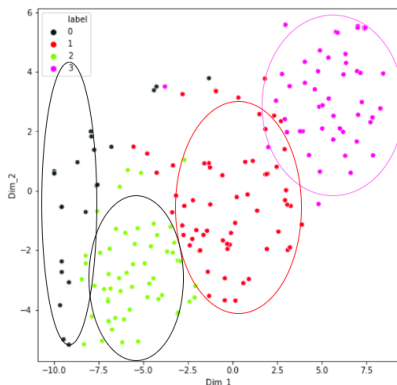
1 - On x-vectors dataset: select K



**Figure 4:** The squared error (Cost) for different values of K (from 1 to 10)

### 3.1.1 Applying K-means on x-vectors data set

#### 2 - Cluster the dataset using the K-means model



**Figure 5:** t-SNE visualization of data grouped in four clusters defined by the K-means algorithm from x-vectors development dataset. t-SNE method reduces the data into two dimensions. The circles represent the four clusters, labeled between 0 to 3.

### 3.1.1 Applying K-means on x-vectors data set

Categories	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Number of files
audiobooks	66.7 %	8.3 %	0	25 %	12
broadcast interviews	0	0	0	100 %	12
child	0	47.83 %	52.17 %	0	23
clinical	0	4.2 %	95.83 %	0	24
court	33.33 %	0	0	66.66 %	12
maptask	13 %	0	0	87 %	23
meeting	0	57.14 %	28.57 %	14.28 %	14
restaurant	0	75 %	25 %	0	12
socio field	25 %	58.33 %	0	16.67 %	12
socio lab	12.5 %	56.25 %	18.75 %	12.5 %	16
webvideo	6.25 %	53.125 %	34.375 %	12.50 %	32
<b>TOTAL</b>	<b>22</b>	<b>63</b>	<b>56</b>	<b>51</b>	<b>192</b>

**Table 3:** Table of the percentage of files by category in each cluster defined by the K-means model on x-vectors development dataset.

## 3.1.1 Applying K-means on x-vectors data set

### 3 - Evaluate the model

The model is evaluate with two scores type:

- **Completeness Score** is satisfied if all data points in a class are grouped in the same cluster [Pedregosa et al. 2011]. This measure is not dependent on absolute label values, i.e. a change in label values will not affect the completeness score.
- **Homogeneity Score** is an indicator to check whether a cluster contains only samples from a single class.

Both Completeness and Homogeneity scores are in the range  $[0,1]$ , the bigger the value the better the model.

### 3.1.1 Applying K-means on x-vectors data set

K value	Homogeneity Score	Completeness Score
2	0.12	0.41
3	0.20	0.41
4	0.30	0.53
5	0.35	0.52
6	0.33	0.45
7	0.37	0.46
8	0.42	0.51

**Table 4: Homogeneity Score and Completeness Score** of K-means models with different k values on x-vectors development dataset.



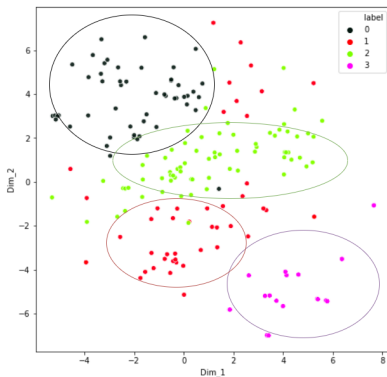
### 3.1.2 Applying diarization on the new classification

Clusters	DER	JER	Threshold
0	6.33	42.23	- 0.6
1	34.30	63.30	- 0.2
2	31.85	63.13	- 0.5
3	10.53	38.19	- 0.4
<b>TOTAL</b>	<b>20.75</b>	<b>51.63</b>	-

**Table 5:** DER and JER scores of each K-means clusters on the 192 files (development dataset).

### 3.1.3 Applying K-means on i-vectors data set

We applied the K-means on the i-vectors data set



**Figure 6:** t-SNE visualization of data grouped in four clusters defined by the K-means algorithm from i-vectors development dataset.

### 3.1.3 Applying K-means on i-vectors data set

Categories	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Number of files
audiobooks	8.33 %	91.67 %	0	0	12
broadcast interviews	100 %	0	0	0	12
child	0	0	100 %	0	23
clinical	95.83 %	0	4.17 %	0	24
court	0	100 %	0	0	12
maptask	0	0	0	100 %	23
meeting	71.42 %	0	0	28.57 %	14
restaurant	0	75 %	25 %	0	12
socio field	0	0	0	100 %	12
socio lab	100 %	0	0	0	16
webvideo	3.13 %	6.25 %	90.62 %	0	32
<b>TOTAL</b>	<b>51</b>	<b>46</b>	<b>72</b>	<b>23</b>	<b>192</b>

**Table 6:** Table of the percentage of files by category in each cluster defined by the K-means model on i-vectors development dataset. The last line of the table refers to the number of files in each cluster.

### 3.1.3 Applying K-means on i-vectors data set

#### Evaluate the model

The final step is to evaluate the model. Table 7 shows the Homogeneity and Completeness scores for a k-value between 2 to 8.

K value	Homogeneity Score	Completeness Score
2	0.16	1
3	0.35	0.93
4	0.48	0.86
5	0.45	0.81
6	0.85	0.85
7	0.63	0.86
8	0.65	0.80

**Table 7: Homogeneity Score and Completeness Score** of k-means model computed with different k values.

### 3.1.4 Applying diarization on the new classification

The K-means algorithm clusters the files in 4 clusters. We applied the diarization on each cluster with the new classification of files.

Clusters	DER	JER	Threshold
0	21.09	41.97	- 0.6
1	9.77	47.88	- 0.4
2	41.13	71.31	- 0.3
3	7.62	14.21	- 0.4
<b>TOTAL</b>	<b>19.9</b>	<b>43.84</b>	-

**Table 8:** DER and JER scores of each K-means clusters on the 192 files (development dataset).

## 3.2 Supervised method: Classification

*A method which depends on **labeled input data**.*

K-Nearest Neighbors (KNN):

- Algorithm to cluster large amount of data into smaller groups together which have similar properties.
- **Principle:** calculate the distance between each point (represented in an n-dimensional space) and the centroid of each cluster. Each point will be assigned to the cluster for which the distance is the smallest.
- **2 important parameters:** k-value (k number of cluster) + distance metric (Euclidean distance).
- **Advantages:** Easy to implement / Fast training / Robustness.
- **Disadvantages:** Biased k value / High complexity.

## 3.2 Supervised method: Classification

### Application of the KNN algorithm on the DIHARD II challenge dataset

- 2 types of data: x-vectors and i-vectors splitted into development and evaluation groups.

	<b>data</b>	<b>label</b>	<b>files</b>
0	[-0.8711261, -0.01523707, -1.312261, -1.124552...	1	DH_0001
1	[1.208957, -0.6493585, -1.5397, 0.01195502, -2...	1	DH_0002
2	[2.979908, 0.7329904, -1.914249, -0.07597715, ...	1	DH_0003
3	[-0.5403174, 0.5555854, -1.923427, -0.08261859...	1	DH_0004
4	[-0.08468243, 0.8757551, -1.595554, 0.2286839,...	1	DH_0005

**Figure 7:** Data from x-vectors. Each data is represented by a vector, a label, and the name of the file.

## 3.2 Supervised method: Classification

### Application of the KNN algorithm on the DIHARD II challenge dataset

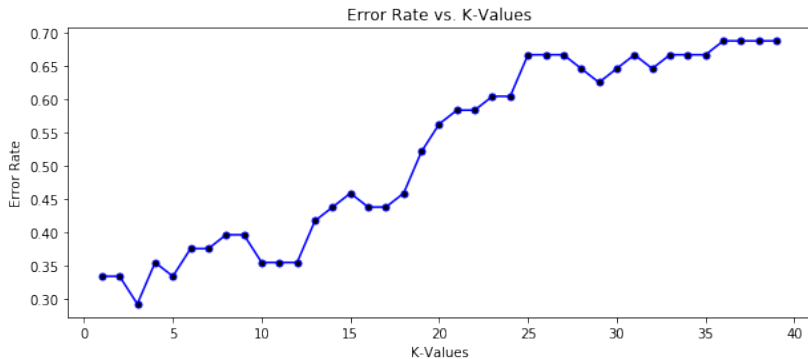
Construction of the KNN model in **three steps**:

1. Train and test the x-vectors development data on several KNN model.
2. Create the model with the previously set k-value.
3. Evaluate the model.



## 3.2 Supervised method: Classification

On x-vectors dataset: first step



**Figure 8:** Error rate according to the k value.

## 3.2 Supervised method: Classification

On x-vectors dataset: second step

	1	2	3	4	5	6	7	8	9	10	11
1	11	0	0	0	1	0	0	0	0	0	0
2	0	10	0	0	2	0	0	0	0	0	0
3	0	0	21	2	0	0	0	0	0	0	0
4	0	0	1	23	0	0	0	0	0	0	0
5	0	0	0	0	12	0	0	0	0	0	0
6	0	0	0	0	0	23	0	0	0	0	0
7	0	0	0	2	0	0	6	2	3	1	0
8	0	0	0	0	0	0	0	12	0	0	0
9	0	0	1	0	1	0	0	0	9	1	0
10	0	0	1	2	0	0	5	0	0	8	0
11	0	1	1	5	0	0	1	6	0	0	18

**Figure 9:** Confusion matrix computed on the KNN model on the 192 files (development dataset) from x-vectors - accuracy of 79.6 %.

## 3.2 Supervised method: Classification

On x-vectors dataset: third step

	1	2	3	4	5	6	7	8	9	10	11
1	3	1	0	1	3	2	0	0	0	0	2
2	0	4	0	0	4	1	1	0	2	0	0
3	0	0	17	2	0	0	0	2	0	0	2
4	0	0	0	24	0	0	0	0	0	0	0
5	0	0	0	0	11	0	0	0	0	1	0
6	0	0	0	0	0	19	0	0	0	0	0
7	0	0	2	3	0	0	2	3	1	0	0
8	0	0	0	0	0	0	0	12	0	0	0
9	0	1	1	6	0	0	8	2	2	0	2
10	0	0	0	0	1	0	6	0	0	4	1
11	0	0	6	4	0	0	1	9	2	0	13

**Figure 10:** Confusion matrix computed on the KNN model on the 194 files (evaluation dataset) from x-vectors - accuracy of 57.2 %.

Example of a new classification

*File originally categorize as*  
*webvideo:* **DH\_0008**

*File from child category:*  
**DH\_0034**

## 3.2 Supervised method: Classification

### Applying diarization with the new classification

Categories	DER	JER	Threshold
audiobooks	0.0	0.0	- 2.0
broadcast interviews	7.14	46.90	- 0.7
child	33.92	64.12	- 0.4
clinical	22.51	42.53	- 0.4
court	12.23	45.60	- 0.2
maptask	8.46	14.12	- 0.4
meeting	28.38	55.65	- 0.5
restaurant	47.00	75.27	- 0.3
socio field	21.23	48.03	- 0.5
socio lab	10.73	33.42	- 1.0
webvideo	47.08	82.56	- 0.7
<b>TOTAL</b>	<b>21.69</b>	<b>46.2</b>	-




**Figure 11:** Table which contains DER and JER of each category from the development dataset(specific threshold) with the new classification.

## 4. Conclusion

- Several issues identified from the DIHARD II challenge dataset [Ryant et al. 2019]: diarization performance is different for different categories of audio files when computed with a global threshold - optimizing the threshold regarding the categories gives better diarization performances.
- t-SNE visualization of the speech embeddings shows that some of the audio categories are nicely clustered and some of those clusters are somewhat closer - find a new classification can help.
- By applying two classification methods, we got better diarization performances, especially with the K-means method. This proves that find new patterns, correlations between files, and therefore a new classification decrease the error rate of diarization.

## 5. Future prospects

- Investigate other classification methods (deep neural network, ...).
- Work on more diverse data with larger categories.
- Create a platform to record a speech and which directly applies the diarization on it.

-  PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.
-  RYANT, N. et al. First dihard challenge evaluation plan. Tech. Rep., 2018.[Online]. Available: <https://zenodo.org/record/1199638>, 2018.
-  RYANT, N. et al. Second dihard challenge evaluation plan. *Linguistic Data Consortium, Tech. Rep*, 2019.

Thank you  
Questions?