Speaker Diarization

4th June 2024

Yannis TEVISSEN

IMA4511 - Pattern Recognition and Biometrics

Who am I ?

- Graduated from Engineering School Télécom SudParis in 2020
- Co-founder of deeptech startup VocaCoach
- PhD Graduate from Institut Polytechnique de Paris

  Supervisors : Jérôme Boudy, Gérard Chollet, Frédéric Petitpont

- Head of Science at Moments Lab

Yannis **Tevissen**

yannis.tevissen@momentslab.com

https://yannistevissen.fr

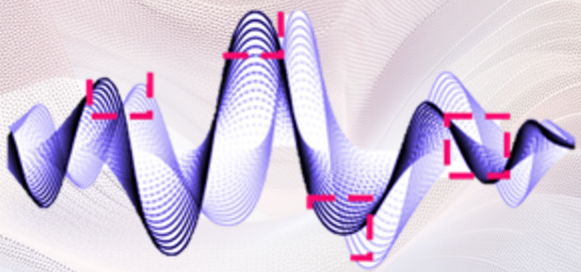
Moments Lab

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

Plan

- Speaker Diarization
- Speaker Diarization has severe Robustness issues
- State-of-the-art algorithms to solve these issues
  - Example of SOTA Diarization in the media industry
  - Speaker Diarization Research in France (Industry and Academics)
  - Media and Healthcare usecases



What is speaker diarization

*“**Speaker diarization** is a task to **label audio or video recordings** with classes that correspond to **speaker identity,** or in short, a task to identify **"who spoke when".**”*

➔ comes from the english ***diary***

History of research in Speaker Diarization

Initially diarization was seen as a component of **Automatic Speech Recognition (ASR)** but over the past years, it has become a field of research of its own.

The **National Institute of Standards and Technology (NIST)** defined the bases of speaker diarization for its Rich Transcription evaluations.

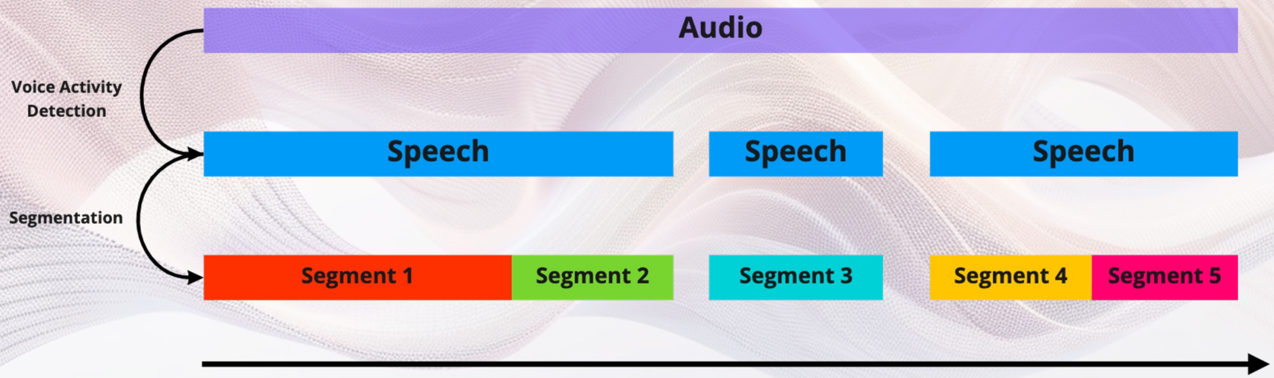
There have been a renewed interest about diarization over the past decade with the new **neural-based methods** made possible by the development of heavy GPU computing.

Segmentation

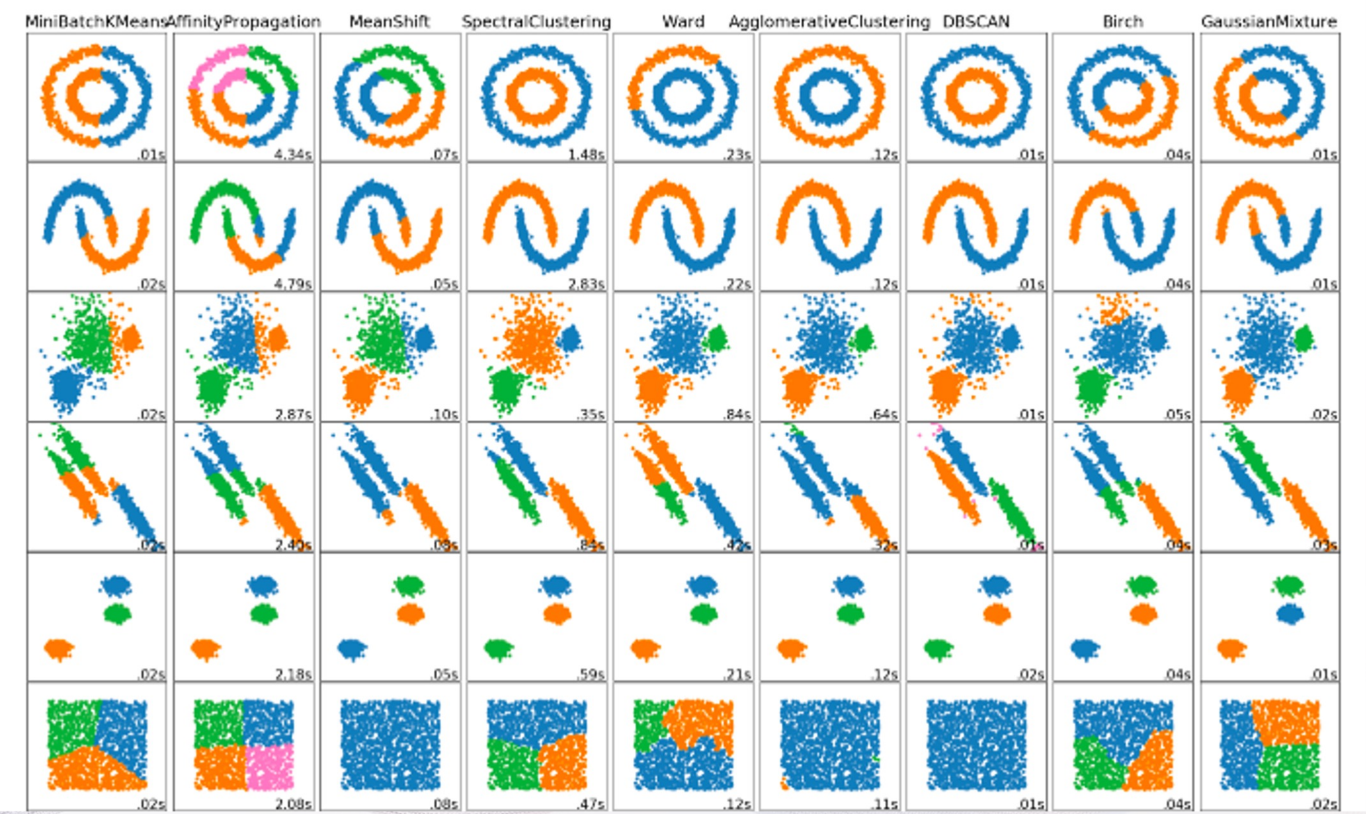
Segmentation is the task of splitting an audio into homogeneous speech segments.

An **homogeneous speech segment** is a segment that contains only one speaker voice.

Prior to this task we often perform pre-processing such as **voice activity detection**.



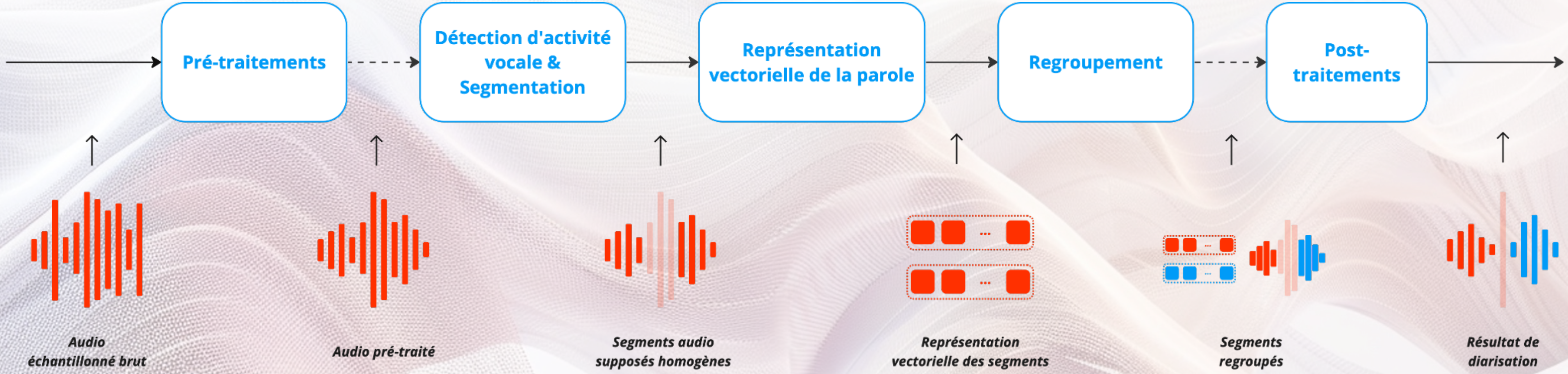
Clustering



MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture

Clustering

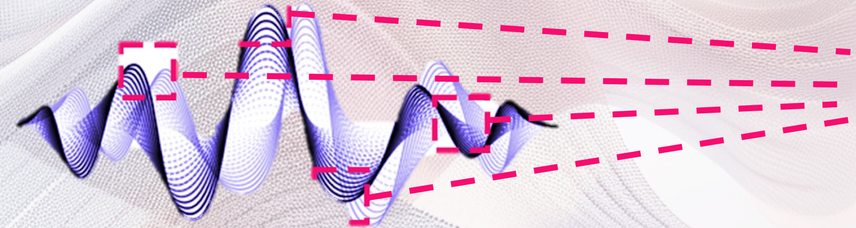
In the case of speaker diarization, we talk about **segment clustering**, with the objective of : 1 cluster = 1 speaker.


Pré-traitements
Détection d'activité vocale & Segmentation
Représentation vectorielle de la parole
Regroupement
Post-traitements
Audio échantillonné brut
Audio pré-traité
Segments audio supposés homogènes
Représentation vectorielle des segments
Segments regroupés
Résultat de diarisation

Speaker diarization and ASR

Diarization is the entry point of every speech to text engine. Transcription technologies have become so mature that they achieve outstanding results when they are presented to the right data.

The objective of speaker diarization is to reproduce the best possible conditions for the transcription algorithms to always work at their best.

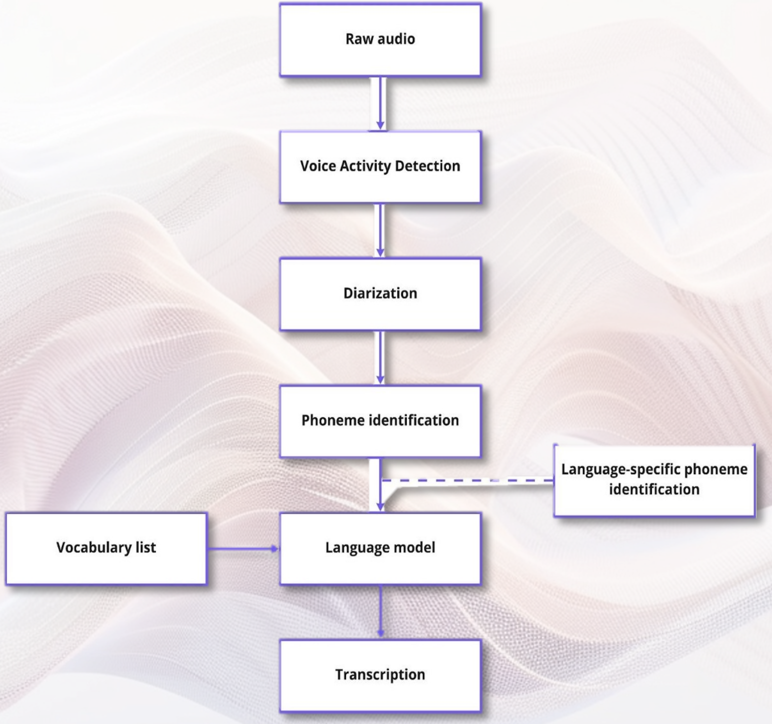


Speaker 1 : 00:01:03 -> 00:01:12
"I have a dream that one day this nation will rise up …"

Speaker diarization and ASR

Example of state-of-the-art Diarization + ASR :

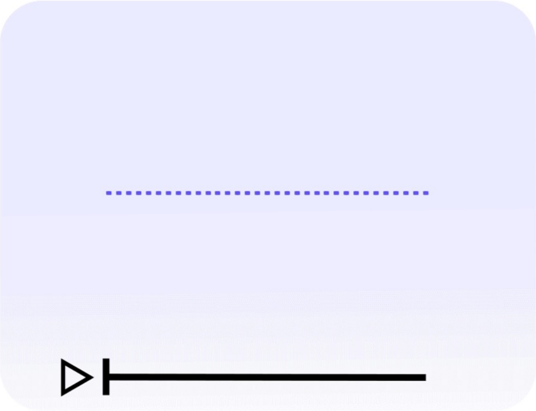
Diarization + Whisper + BERT



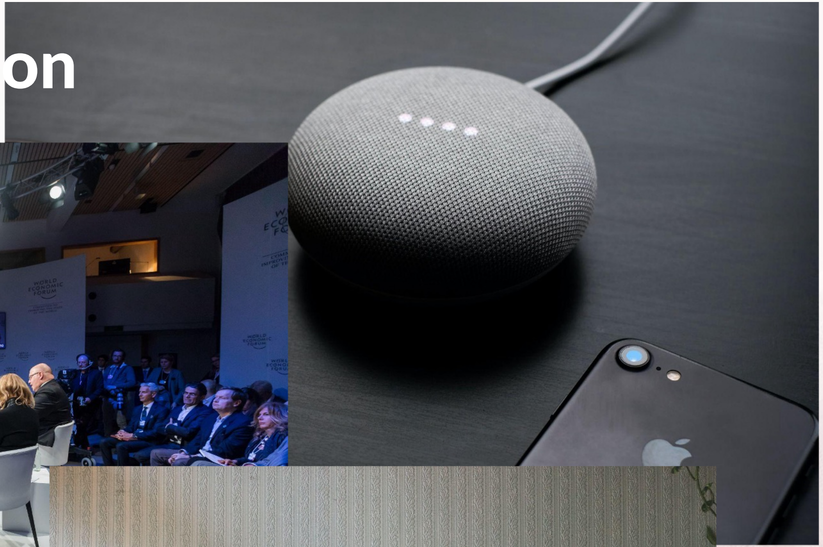
https://github.com/m-bain/whisperX

Speaker diarization and ASR

Speaker diarization can help **improving ASR** and reducing Word Error Rates.



Use cases for speaker diarization



Use case : Smart Homes

- Counting the **number of speakers**,
- Adapting **voice interfaces** in case of multiple speakers,
- But also :
  - Dynamically adapting temperature,
  - Acoustic Person Tracking,
  - Smart decisions in case of emergency.

e-ViTA
EU-JAPAN VIRTUAL COACH FOR SMART AGEING

Use case : Media industry

- **MediaHub** for broadcasted live and archive content
  - Indexing content thanks to multimodal AI,
  - Efficiently searching in thousands of media hours,
  - Re-selling relevant media.

Use case : Panel and meeting retranscription

- **Faster retranscription** of debates
- **Better accessibility** for hearing impaired people

“hello hello how are you fine thank you what about you perfect let’s start wait a minute please ok no problem thanks”

“ - hello
- *hello how are you*
- *fine thank you what about you*
- *perfect let’s start*
- *wait a minute please*
- *ok no problem*
- *thanks”*

Use case : Defense

- **Phone surveillance**

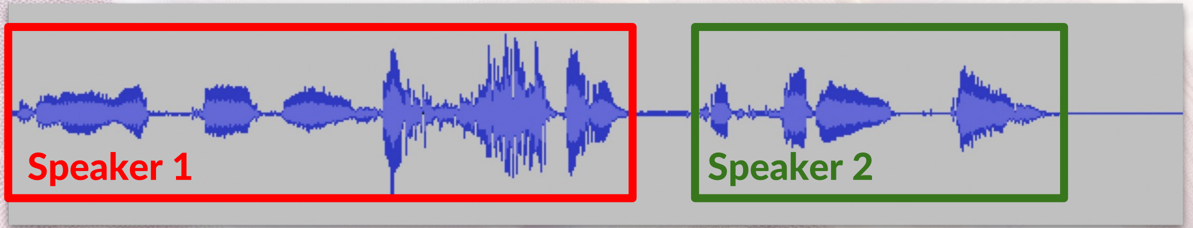
- Separation and Diarization prior to **speaker identification**

Robustness of Speaker Diarization

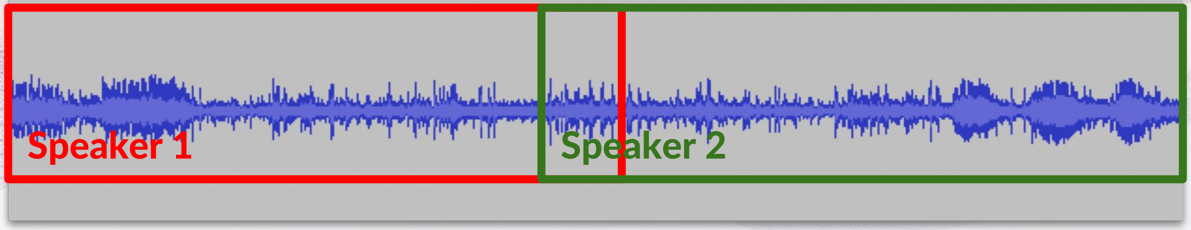
PhD Subject :

Audiovisual diarization : Reaching model robustness in the wild

Theory :



Real world :



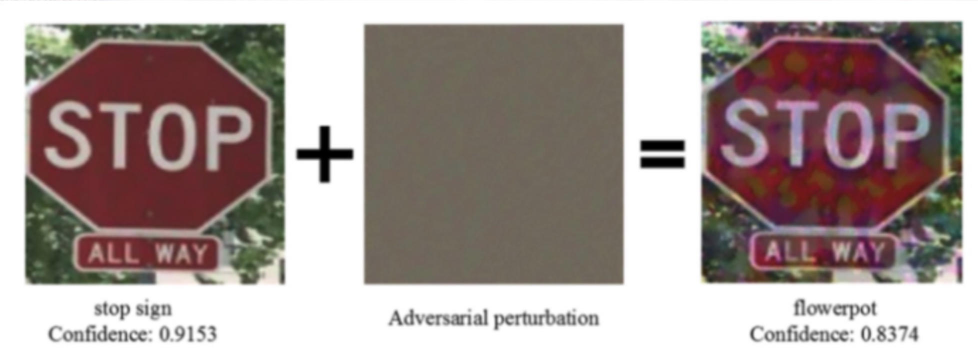
Algorithmic robustness

Robustness, according to IEEE : *The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions*

- Adversarial robustness
- Noise robustness
- Biases handling
- Overlapping speech handling

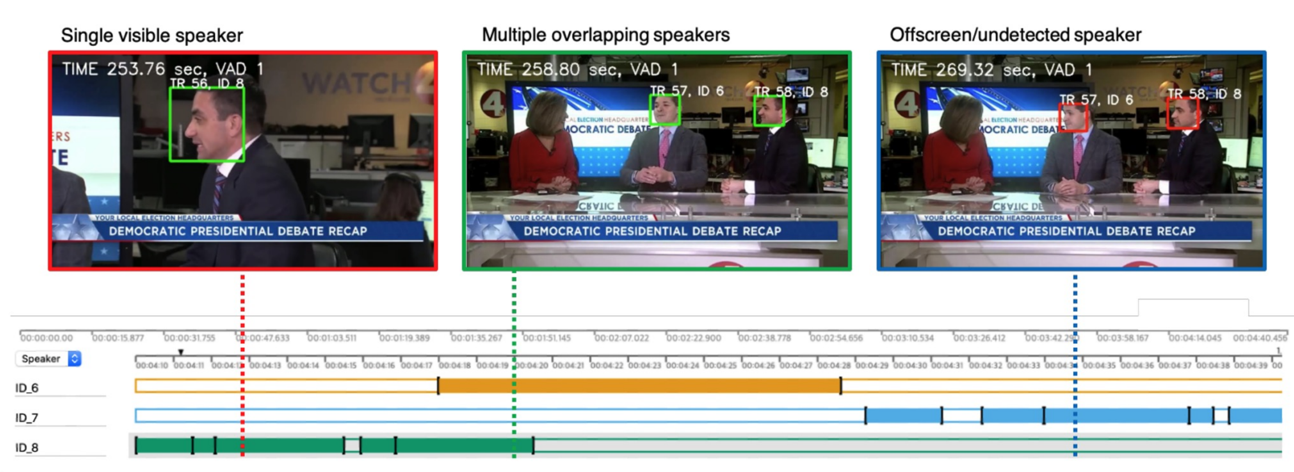
Adversarial robustness

Adversarial robustness in **Machine Learning & Deep Learning**



Robustness comes with performances

We talk about "**in the wild**" performances.



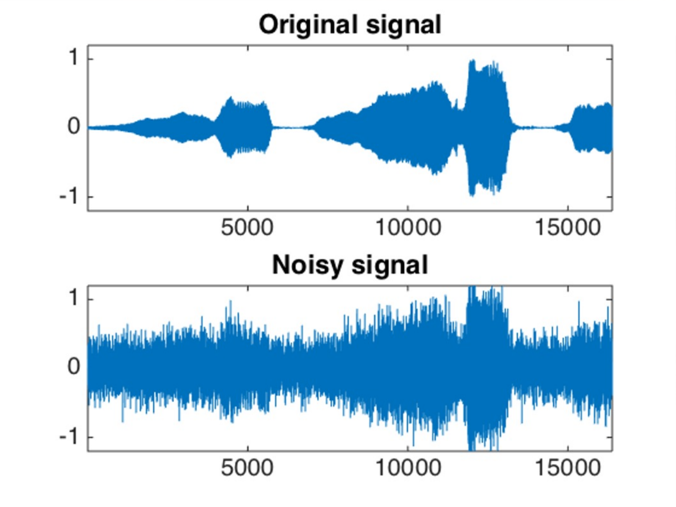
J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "**Spot the conversation: Speaker diarisation in the wild**" Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2020-Octob, pp. 299–303, 2020, doi: 10.21437/Interspeech.2020-2337.

Signal-to-Noise Ratio

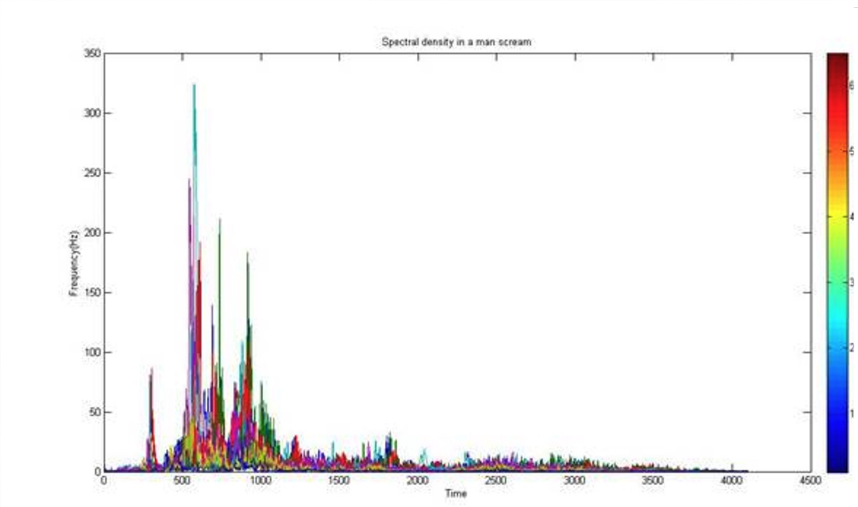
$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}$$

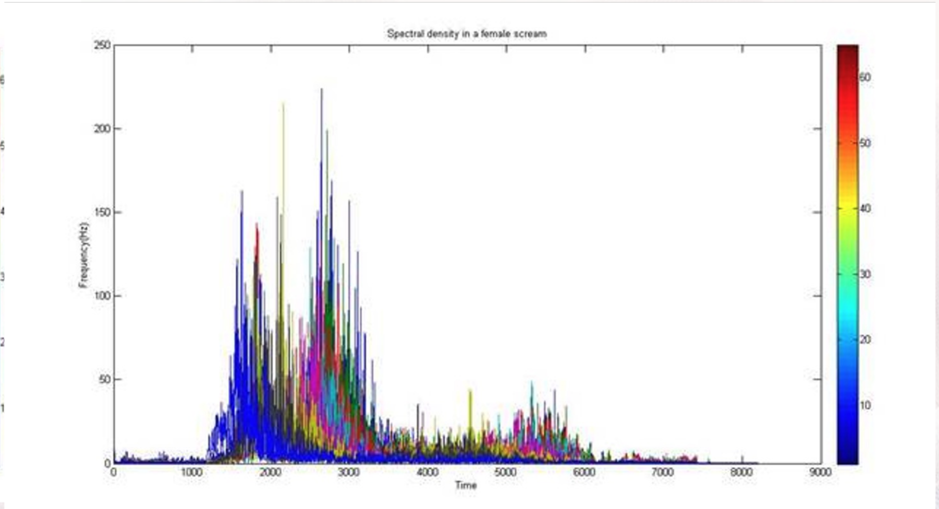
$$\mathrm{SNR_{dB}} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right)$$

- Difficulty for real-world audio content
- SNR cannot be calculated afterwards
- Good model for white noise studies



Biases in speech technologies





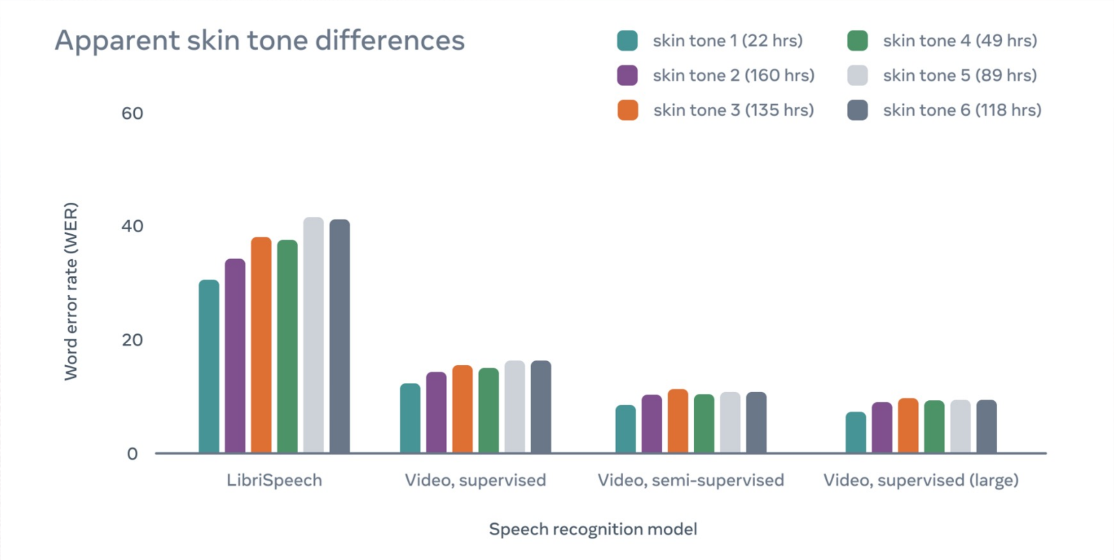
Gender biases

Age biases

Disability biases

Accent biases

Biases in speech technologies



Overlapping Speech

Overlapping speech is when **two or more people are speaking at the same time.**

Very frequent in everyday conversations and in TV debates.

Number of speakers

The more you have speaker, the more they may have similar voices.

Number of speaker as an input or as **a constraint** for some algorithms.

Seeking for solution

Research in speaker diarization is now focused on making it more robust in various contexts:

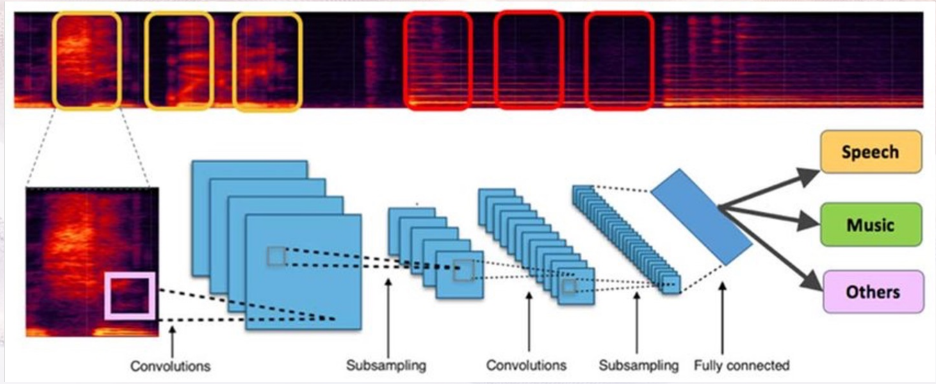
- In-house recording with **multi-channels** and multi-microphones setups
- **TV content**
- Faster diarization and especially clustering techniques for **online processing**
- Multimodal diarization for **damaged audio cases**

Robust Voice Activity Detection

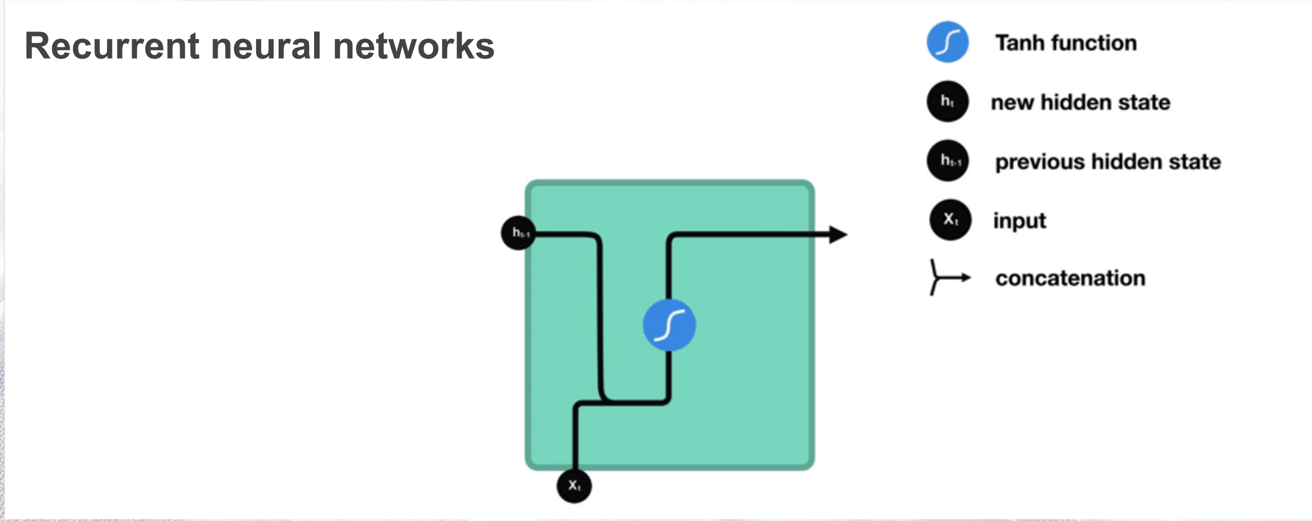
C-RNN are more robust voice activity detector

- **Different strategies** exist
  - Two classes classifiers : Speech / Non Speech
  - Multi classes classifiers : Speech / Animal sounds / Car sounds / …

Example of 3 classes audio classifier :



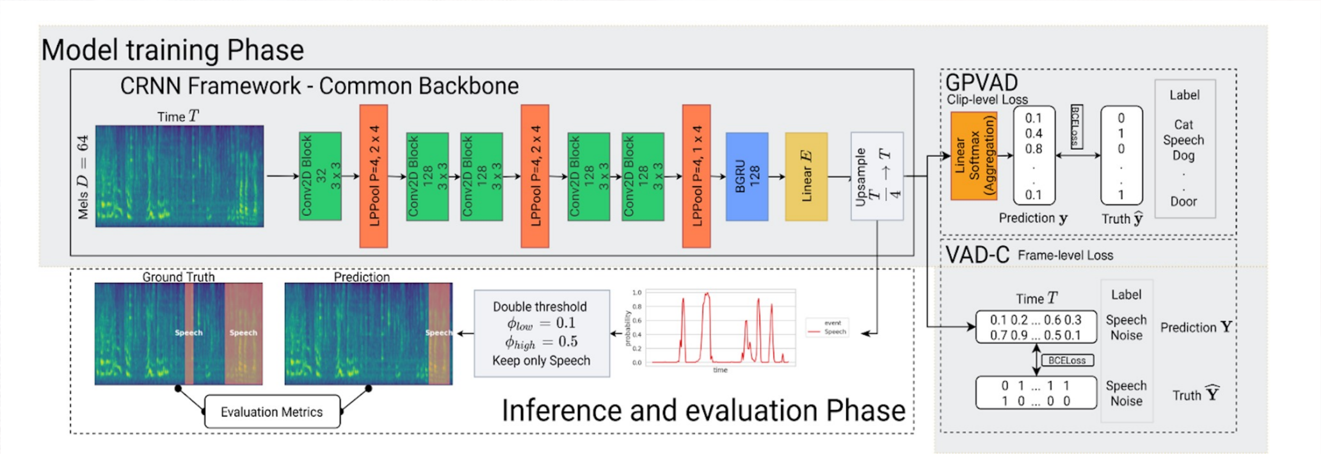
RNNs



new weight = weight - learning rate*gradient

Robust Voice Activity Detection

Weakly supervised sound event detection : 517 classes classifier

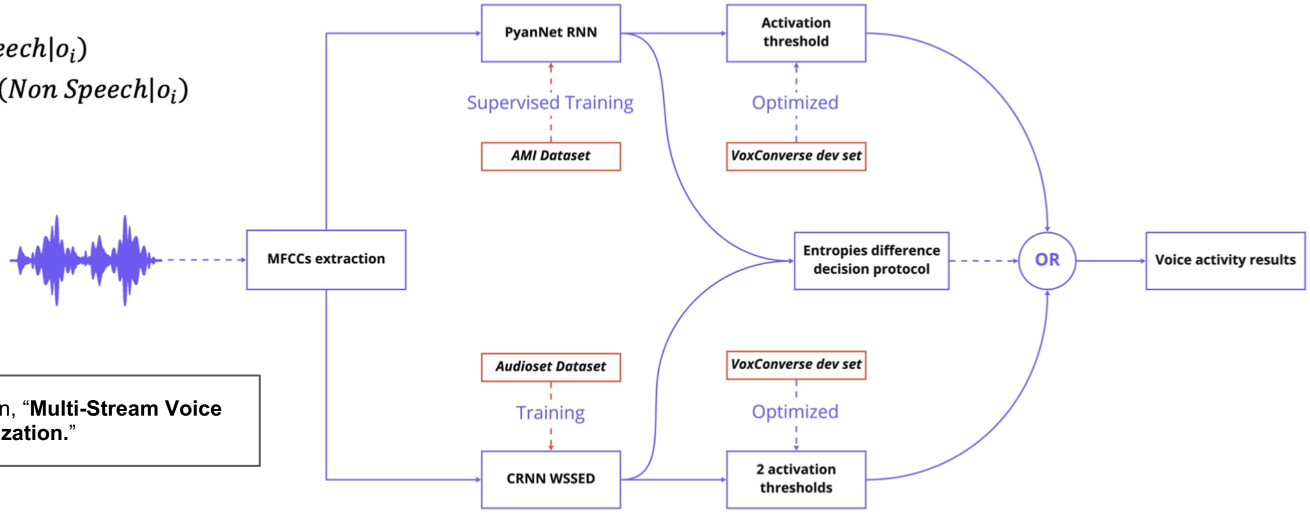


H. Dinkel, Y. Chen, M. Wu, and K. Yu, "**Voice activity detection in the wild via weakly supervised sound event detection**," 2020, [Online]. Available: http://arxiv.org/abs/2003.12222.

Entropy based method selection

Multi-stream Voice Activity Detection for Speaker Diarization

$$h_i = -P(Speech|o_i)\, log_2\, P(Speech|o_i) \\ -P(Non\ Speech|o_i)\, log_2\, P(Non\ Speech|o_i)$$



Y. Tevissen, J. Boudy, F. Petitpont, G. Guenon, “**Multi-Stream Voice Activity Detection for Robust Speaker Diarization.**”

Robust speech embedding

Speech embedding is a way to represent speech in **a lower dimensional vector space**.

It allows us to define rules in terms of **similarity** through distances between two vectors.

To build **robust embeddings** you need to answer a few questions :

- At what level do you want a vector representation ? Frame, segment, …
- What dimension do you need for your embedding vectors ?
- What features do you use to build these embeddings and your vector space ?

Embeddings : i-vectors

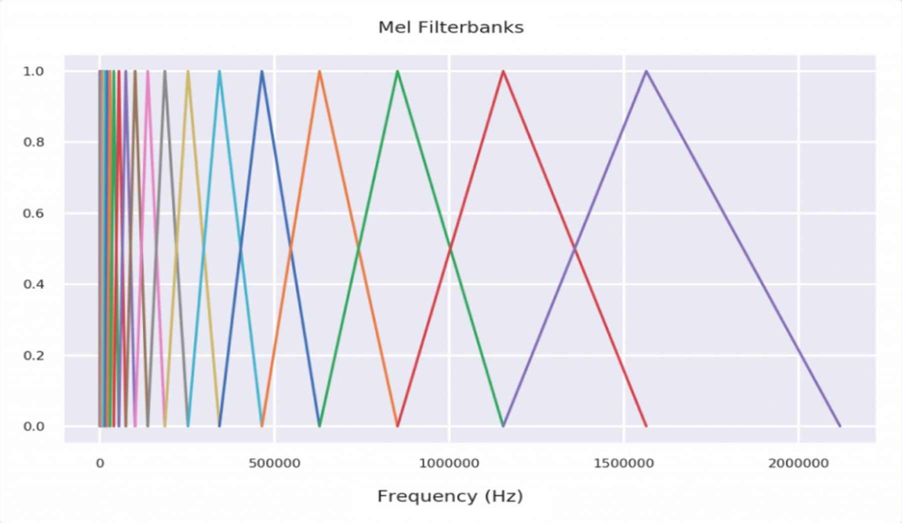
MFCCs = Mel frequency cepstral coefficients

GMM = Gaussian Mixture Models

UBM = Universal Background Model

$$\mathbf{M} = m + Tw$$

M: supervector
m: UBM
T: total variability matrix
w: i-vector



Najim Dehak, Patrick Kenny, Pierre Dumouchel, Reda Dehak, Pierre Ouellet, «**Front-end factor analysis for speaker verification** » in IEEE Transactions on Audio, speech and Language Processing 2011

Embeddings : d-vectors

- MLP Neural-based embedding
- Frame level embedding

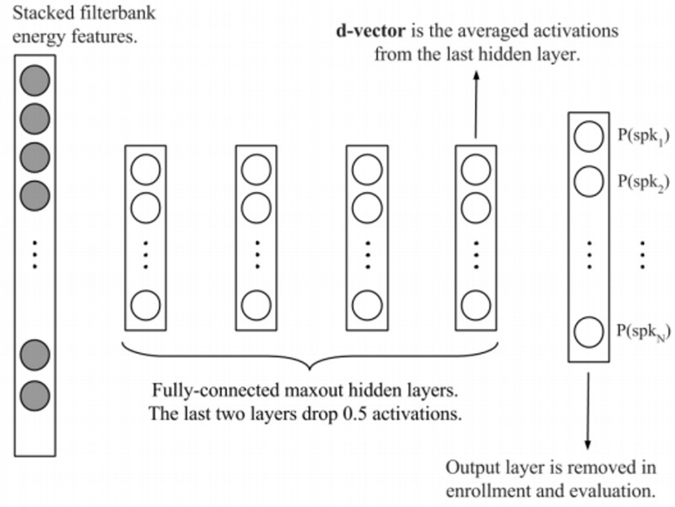
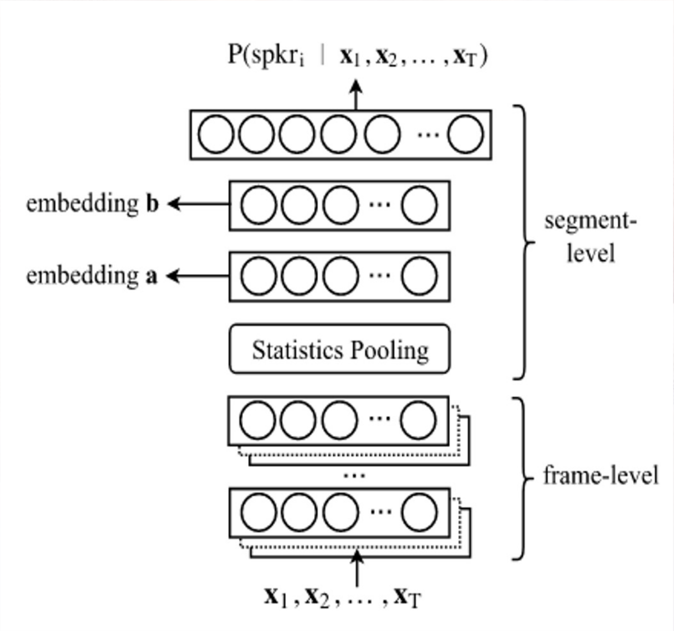


Fig. 1. The background DNN model for speaker verification.

Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "**Deep neural networks for small footprint text-dependent speaker verification**," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 4052– 4056

Embeddings : x-vectors

- Neural-based embedding
- Statistical pooling to go from frame-level to segment level
  - More robust
  - Less adapted to online systems



D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “**X-Vectors: Robust DNN Embeddings for Speaker Recognition**,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2018-April, pp. 5329–5333, 2018, doi: 10.1109/ICASSP.2018.8461375.

Different clustering for different diarization

- Spectral clustering
- K-means
- Hierarchical clustering
- Online UIS-RNN
- Overlap-aware
- Bayesian HMM

```
SPEAKER IN1013 1 37.607 1.639 <NA> <NA> MIO078 <NA> <NA>
SPEAKER IN1013 1 46.490 1.094 <NA> <NA> MIE034 <NA> <NA>
SPEAKER IN1013 1 53.6 5.8 <NA> <NA> MIE034 <NA> <NA>
SPEAKER IN1013 1 54.72 1.623 <NA> <NA> MIO097 <NA> <NA>
SPEAKER IN1013 1 58.76 31.56 <NA> <NA> MIO097 <NA> <NA>
SPEAKER IN1013 1 91.09 37.518 <NA> <NA> MIO097 <NA> <NA>
```

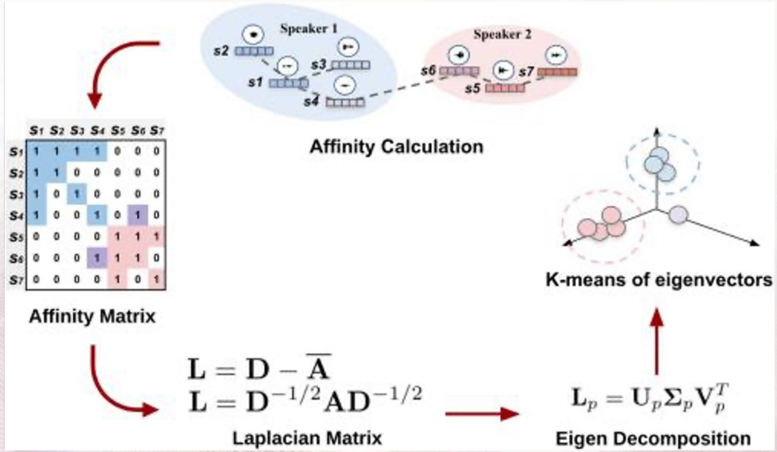
Example of diarization output .rttm file

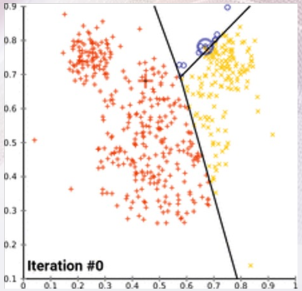
Hierarchical clustering

Top-down vs. Bottom-up clusterings

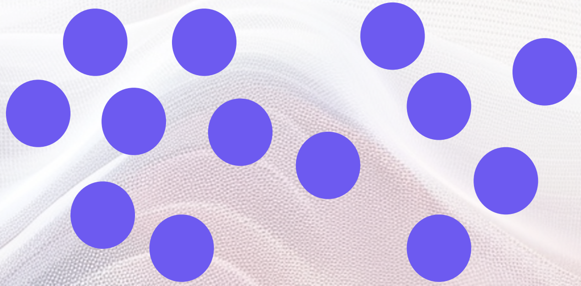
Agglomerative hierarchical clustering (**AHC**) :

- Each segment starts in one cluster
- Clusters are iteratively merged following a **linkage criterion**

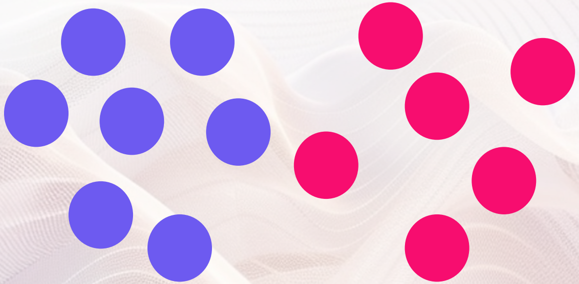




Online clustering



Option 1: one single cluster

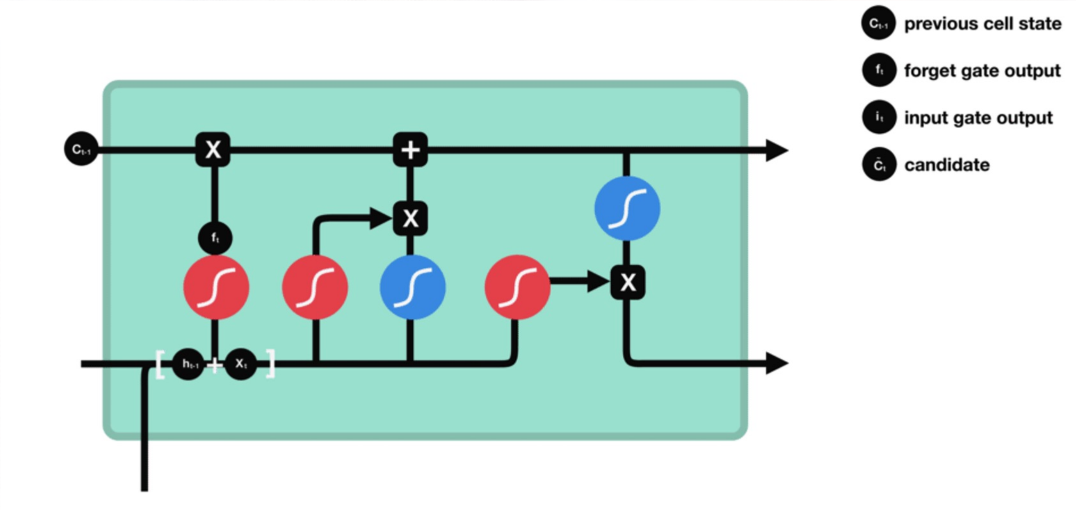


Option 2: two clusters

Online clustering

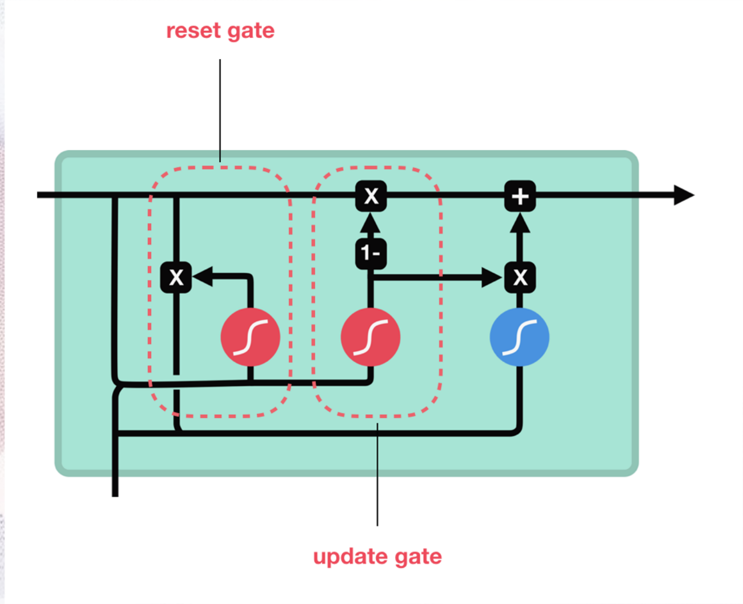
Online clustering : LSTMs

Long Short-Term Memory to solve the gradient vanishing issue



Online clustering : GRUs

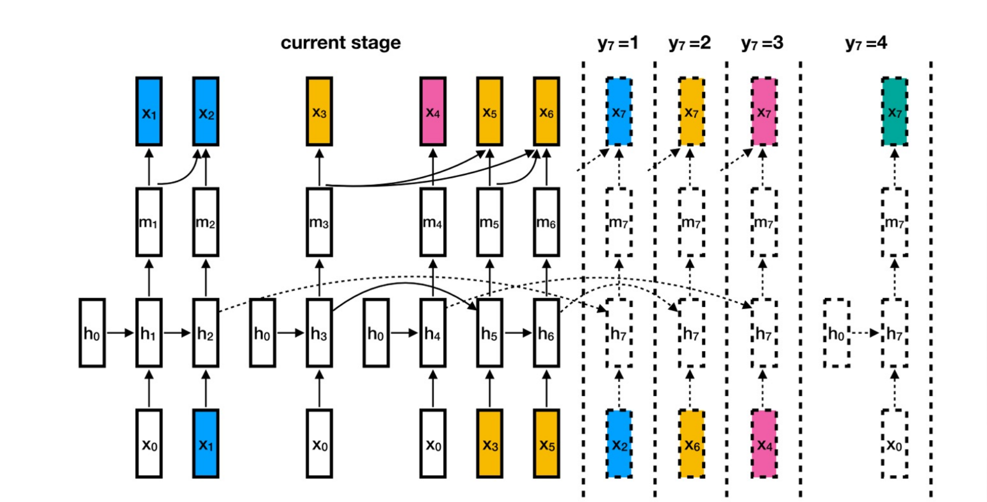
Gated Recurrent Unit



Online clustering : Google way

Only works with frame level embeddings (d-vectors in this case)

Use of Gated Recurrent Units (GRU)

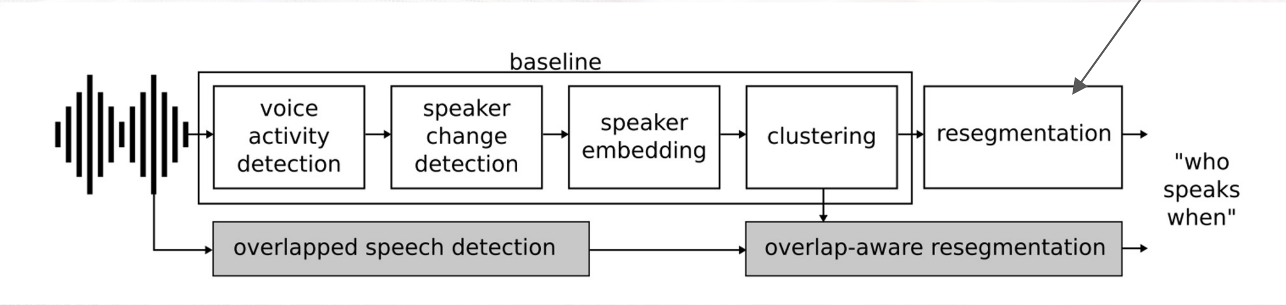


A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "**Fully Supervised Speaker Diarization**," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2019-May, pp. 6301–6305, 2019, doi: 10.1109/ICASSP.2019.8683892.

Overlap-aware clustering

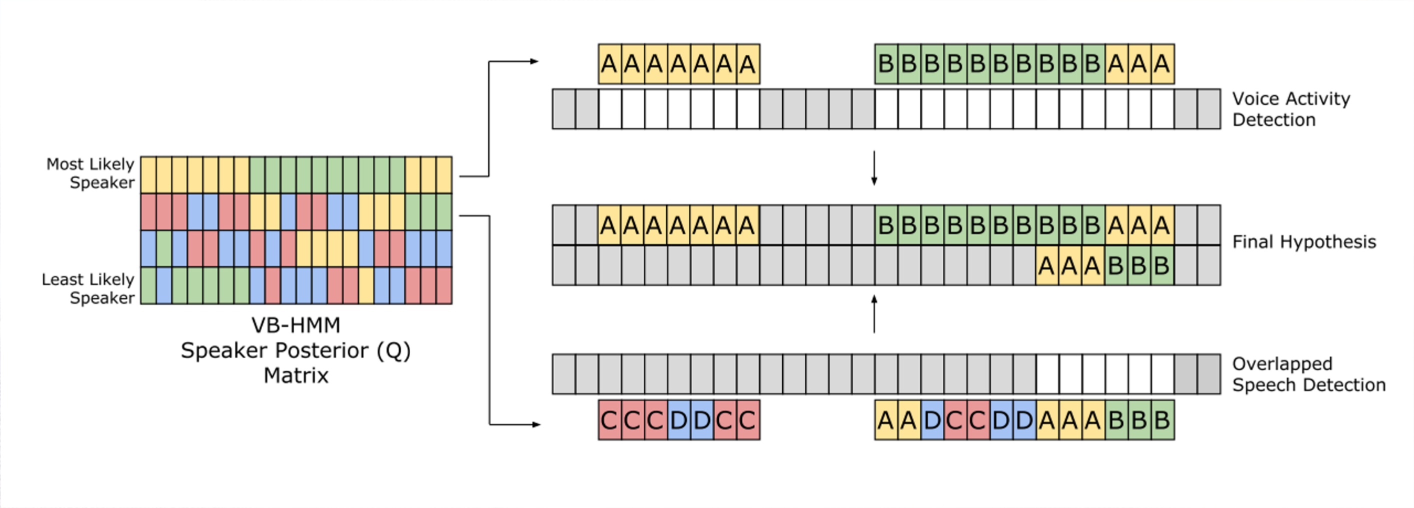
Overlap is hard to define and even harder to process

Viterbi algorithm



L. Bullock, H. Bredin, and L. P. Garcia-Perera, "**Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection**," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 7114–7118, 2020, doi: 10.1109/ICASSP40776.2020.9053096.

Overlap-aware clustering



L. Bullock, H. Bredin, and L. P. Garcia-Perera, "**Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection**," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 7114–7118, 2020, doi: 10.1109/ICASSP40776.2020.9053096.

Multimodality

According to the European Language Resources Association, multimodal technologies refer to "**all technologies combining features extracted from different modalities (text, audio, image, etc.)**."

Multimodal learning ≠ Multimodality

Multimodality : Audio-visual Diarization

- Using visual informations on top of audio
  - Presence of a face
  - Lip’s movements
  - Context of a media

- Works even when audio quality is poor

Some pipelines are entirely open-source:

https://github.com/TaoRuijie/TalkNet-ASD

Transformers

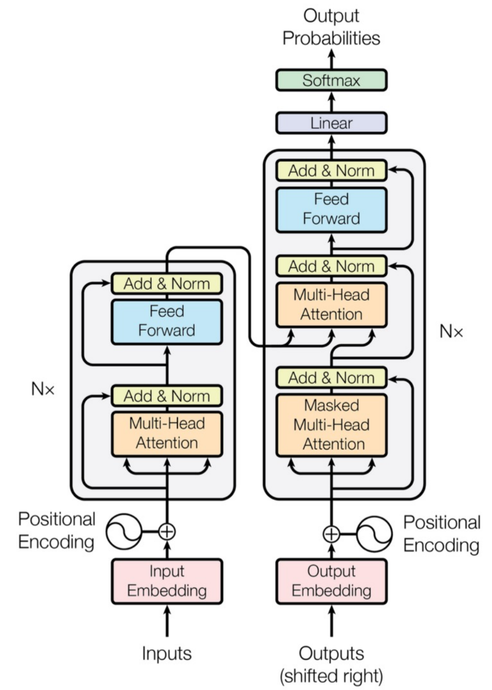


Figure 1: The Transformer - model architecture.

- Sequence to sequence tasks (translation, audio to phonemes, etc.)
- Attention based Encoder / Decoder system

A. Vaswani et al., “**Attention is all you need**,” 2017.

Active Speaker Detection

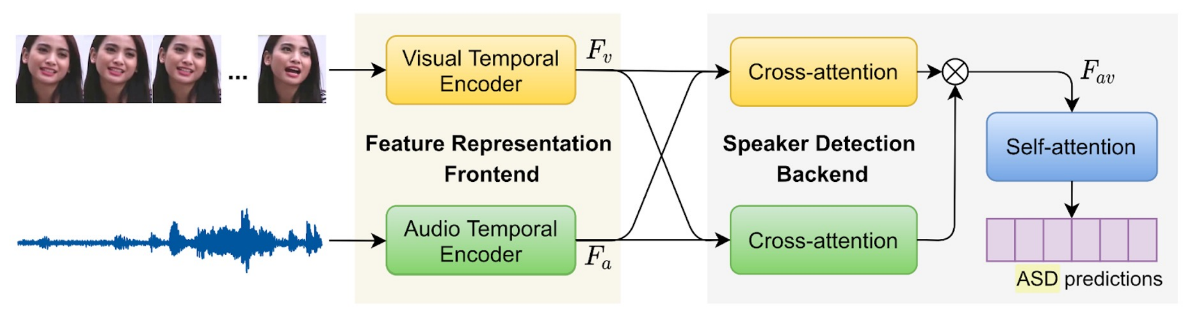


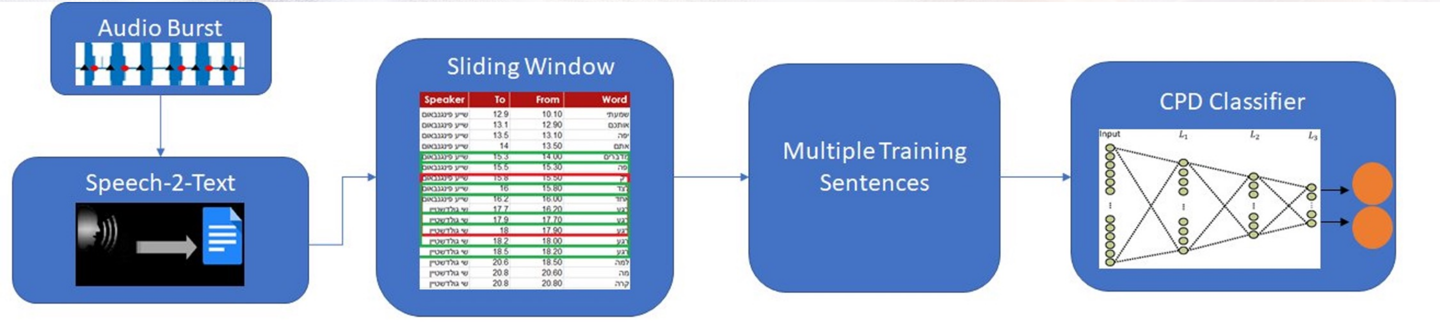
Table 4: Comparison with the state-of-the-art on the AVA-ActiveSpeaker test set in terms of mAP.

Method	mAP (%)
Roth et al. [35]	82.1
Zhang et al. [50]	83.5
Alcazar et al. [5]	86.7
Chung et al. [12]	87.8
TalkNet (proposed)	**90.8**

R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, **Is Someone Speaking?**, vol. 1, no. 1. Association for Computing Machinery, 2021.

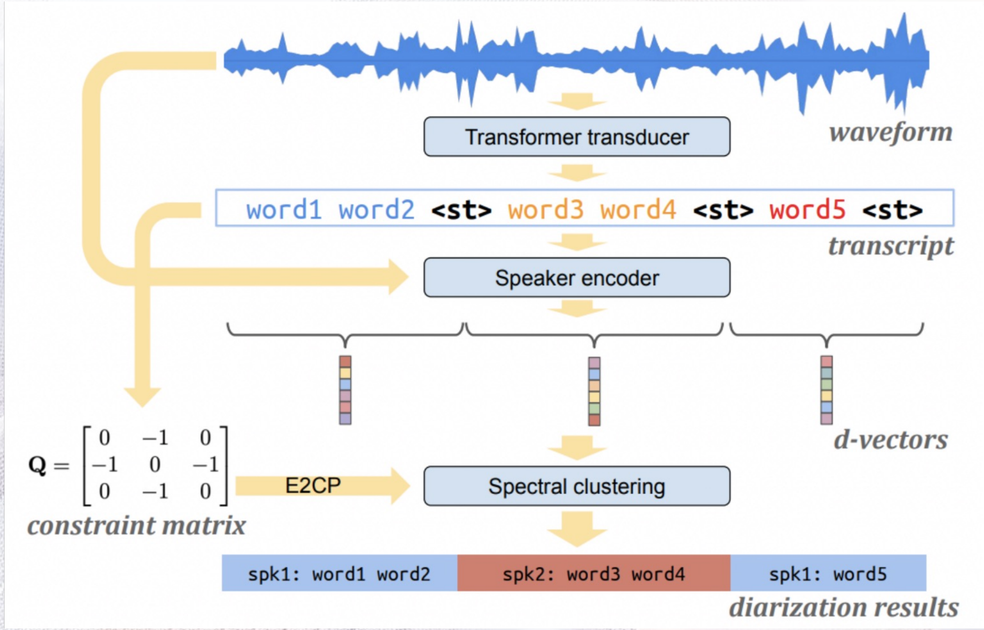
Natural Language Processing for Diarization

Using textual data obtained via ASR to improve speaker diarization



Anidjar, O. & Hajaj, Chen & Dvir, A. & Gilad, I.. (2020). **A Thousand Words are Worth More Than One Recording: NLP Based Speaker Change Point Detection**.

Turn-to-Diarize



W. Xia et al., “**Turn-to-Diarize: Online Speaker Diarization Constrained by Transformer Transducer Speaker Turn Detection**,” no. 2, 2021, [Online]. Available: http://arxiv.org/abs/2109.11641.

Change of paradigm : TS-VAD

Target-Speaker Voice Activity Detection

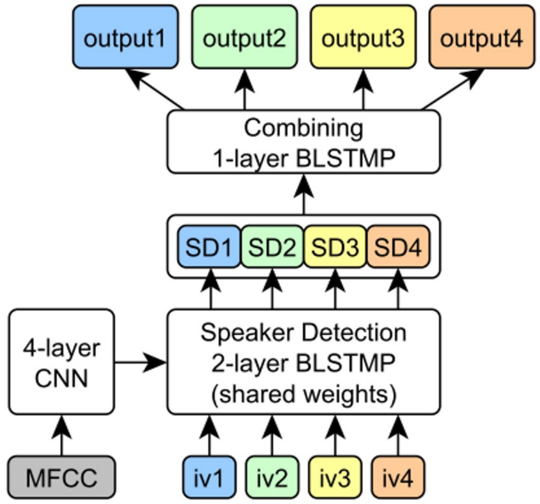


Figure 1: *Single-channel TS-VAD scheme*

	DEV		EVAL	
	DER	JER	DER	JER
x-vectors + AHC	63.42	70.83	68.20	72.54
EEND + WRN x-vectors	52.20	57.42	56.01	61.49
WRN x-vectors + AHC	53.45	56.76	63.79	62.02
WRN x-vectors + SC	47.29	49.03	60.10	57.99
+ TS-VAD-1C (it1)	39.19	40.87	45.01	47.03
+ TS-VAD-1C (it2)	35.80	37.38	39.80	41.79
+ TS-VAD-MC	34.59	36.73	37.57	40.51
Fusion	**32.84**	**36.31**	**36.02**	**40.10**
Fusion*	41.76	44.04	40.71	45.32

Table 2: *Diarization results (* stands for DIHARD II reference)*

I. Medennikov et al., "**Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario**" Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2020-Octob, pp. 274–278, 2020, doi: 10.21437/Interspeech.2020-1602.

Research in France

- LIUM : Sylvain Meignier
- IRIT : Hervé Bredin
- Telecom Paris / Telecom SudParis

Requêtes associées · En progression

1	kaldi	Record
2	google speech to text	Record
3	pyaudioanalysis	Record
4	fully supervised speaker diarization	Record
5	joint speech recognition and speaker diarizatio...	Record

- Open source projects : pyannote, S4D

https://github.com/pyannote/pyannote-audio

pyannote

Use case media & broadcast

TV contents contain two opposite types of media:

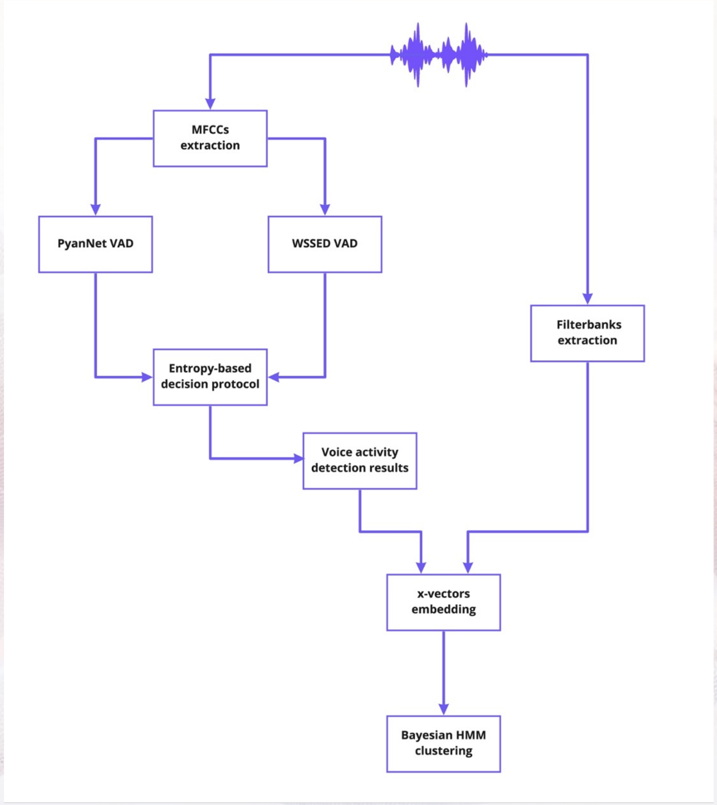
- Media with **good sound quality**
- Media with **noisy sound**

Our Diarization choice

Algorithm 1. Speaker diarization with multi-stream voice activity detection.

Define: W, θ_1, θ_2, θ_3, θ_4
M = MFCC(W)
$V_{PyanNet}$ = PyanNet(M); V_{WSSED} = WSSED(M)
$H_{PyanNet}$ = Entropy($O_{PyanNet}$); H_{WSSED} = Entropy(O_{WSSED});
if $H_{WSSED} - H_{PyanNet} > \theta_1$ **then**
    VAD = Thresholds($V_{PyanNet}$, θ_2)
else
    VAD = Thresholds(V_{WSSED}, θ_3, θ_4)
Diarization = VBx(VAD, W)

Figure 3: *Algorithmic summary of our proposed method. W represents the input audio signal sampled at 16 kHz and θ_1, θ_2, θ_3, θ_4 are optimized on VoxConverse development set.*



Baseline audio approach : Chosen clustering

- Hidden Markov Models
- Ergodic HMM

$$p(s|s') = (1 - P_{\text{loop}})\pi_s + \delta(s = s')P_{\text{loop}}$$

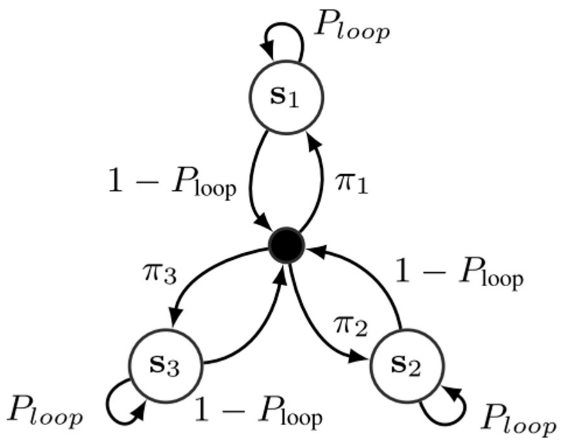


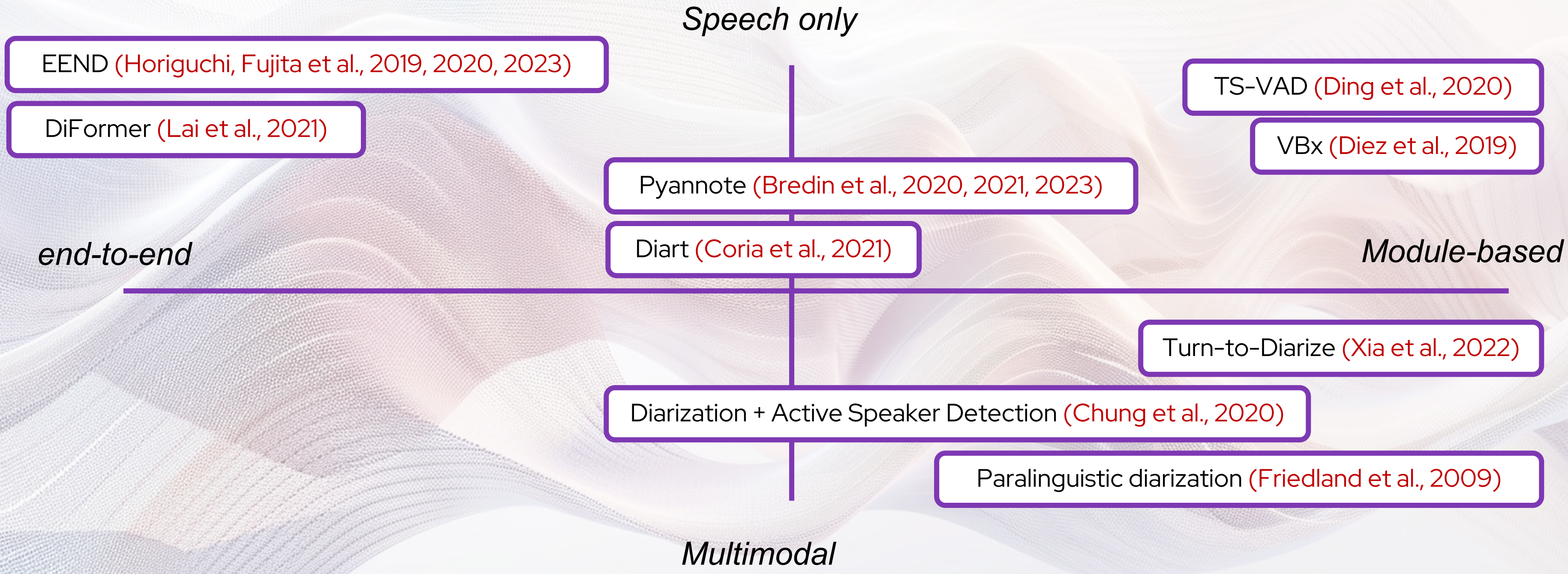
Figure 1: *HMM model for 3 speakers with 1 state per speaker, with a dummy non-emitting (initial) state.*

Baseline audio approach : VBx Diarization

- Handling various audio quality : **Strong Voice Activity Detection**
- Good granularity and robustness : x-vectors embedding
- Overlap-aware : **AHC+HMM clustering**
- Good overall performances

On VoxConverse development set				
VAD Method (+VBx)	**MS**	**FA**	**SC**	**DER**
Energy threshold	9.90	8.27	2.88	21.04
PyanNet RNN	**4.17**	8.94	**2.44**	15.54
WSSED	7.9	**2.13**	2.53	12.56
Multi stream entropy based	5.02	4.14	2.55	**11.69**
Multi stream oracle	5.02	3.50	2.23	10.75

SOTA Speaker Diarization



Use case media & broadcast : engineering

Constraints for production usage :

- Run time
- Long media analysis
- Number of speakers

➔ Necessity to define engineering strategies

Use case healthcare

Goals :

- Energy savings: presence rather than motion detection
- Home care for the elderly:
  - Detect the presence of an elderly visitor:
  - Housekeeper
  - Meal service
  - Factor
  - Family
- Information on social life
- Adapting the distress detection system to the presence of several people

Use case healthcare

- Multi-Stream Voice Activity Detection (MSVAD) (Tevissen, et al. 2022),
- Diart (Coria, et al. 2021).

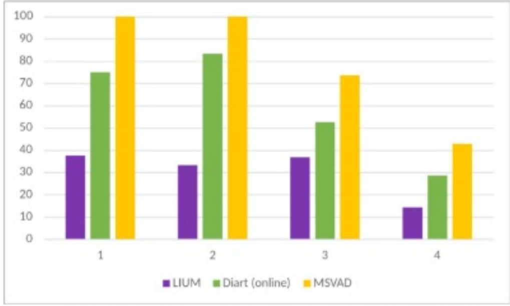


Figure 2. Detailed results of the percentage of correct speaker count for each system depending on the number of active speakers in the recording.

Method	Percentage of correctly labeled recording
LIUM_SpkDiarization	(32.5 ± 2.4) %
Diart (online)	(57.5 ± 3.1) %
MSVAD Diarization	**(77.5 ± 3.6) %**

Contact & Practical Session

Yannis **Tevissen**

yannis.tevissen@momentslab.com

https://yannistevissen.fr

Install Audacity

Make sure you have access to a Google account

Bibliography 1/3

X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," IEEE Trans. Audio, Speech Lang. Process., vol. 20, no. 2, pp. 356–370, 2012, doi: 10.1109/TASL.2011.2125954.

T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A Review of Speaker Diarization: Recent Advances with Deep Learning," 2021, [Online]. Available: http://arxiv.org/abs/2101.09624.

H. Bredin et al., "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 7124–7128, 2020, doi: 10.1109/ICASSP40776.2020.9052974.

M. Diez, L. Burget, F. Landini, and J. Cernocky, "Analysis of speaker diarization based on Bayesian HMM with eigenvoice priors," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 28, no. June, pp. 355–368, 2020, doi: 10.1109/TASLP.2019.2955293.

Y. Tevissen, J. Boudy, F. Petitpont, G. Guenon, T. Sudparis, and I. P. De Paris, "Multi-Stream Voice Activity Detection for Robust Speaker Diarization."

Bibliography 2/3

J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, “Spot the conversation: Speaker diarisation in the wild” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2020-Octob, pp. 299–303, 2020, doi: 10.21437/Interspeech.2020-2337.

H. Dinkel, Y. Chen, M. Wu, and K. Yu, “Voice activity detection in the wild via weakly supervised sound event detection,” 2020, [Online]. Available: http://arxiv.org/abs/2003.12222.

Najim Dehak, Patrick Kenny, Pierre Dumouchel, Reda Dehak, Pierre Ouellet, «Front-end factor analysis for speaker verification » in IEEE Transactions on Audio, speech and Language Processing 2011

Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 4052– 4056

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2018-April, pp. 5329–5333, 2018, doi: 10.1109/ICASSP.2018.8461375.

Bibliography 3/3

A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, “Fully Supervised Speaker Diarization,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2019-May, pp. 6301–6305, 2019, doi: 10.1109/ICASSP.2019.8683892.

L. Bullock, H. Bredin, and L. P. Garcia-Perera, “Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 7114–7118, 2020, doi: 10.1109/ICASSP40776.2020.9053096.

R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, Is Someone Speaking?, vol. 1, no. 1. Association for Computing Machinery, 2021.

Anidjar, O. & Hajaj, Chen & Dvir, A. & Gilad, I.. (2020). A Thousand Words are Worth More Than One Recording: NLP Based Speaker Change Point Detection.

I. Medennikov et al., “Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2020-Octob, pp. 274–278, 2020, doi: 10.21437/Interspeech.2020-1602.