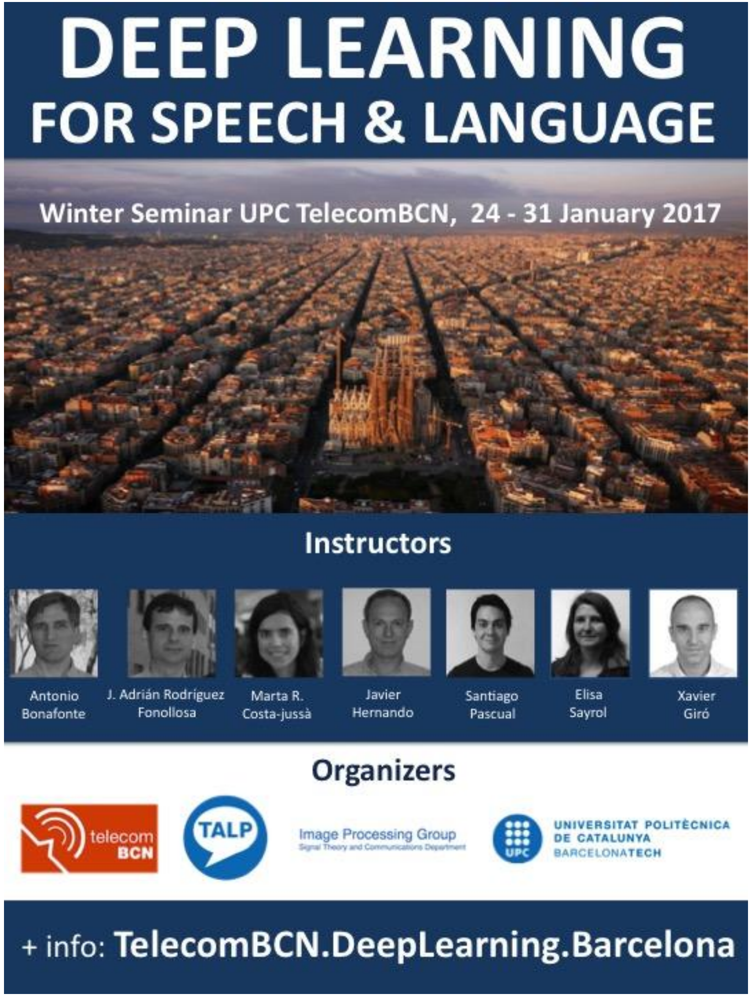


[course site]

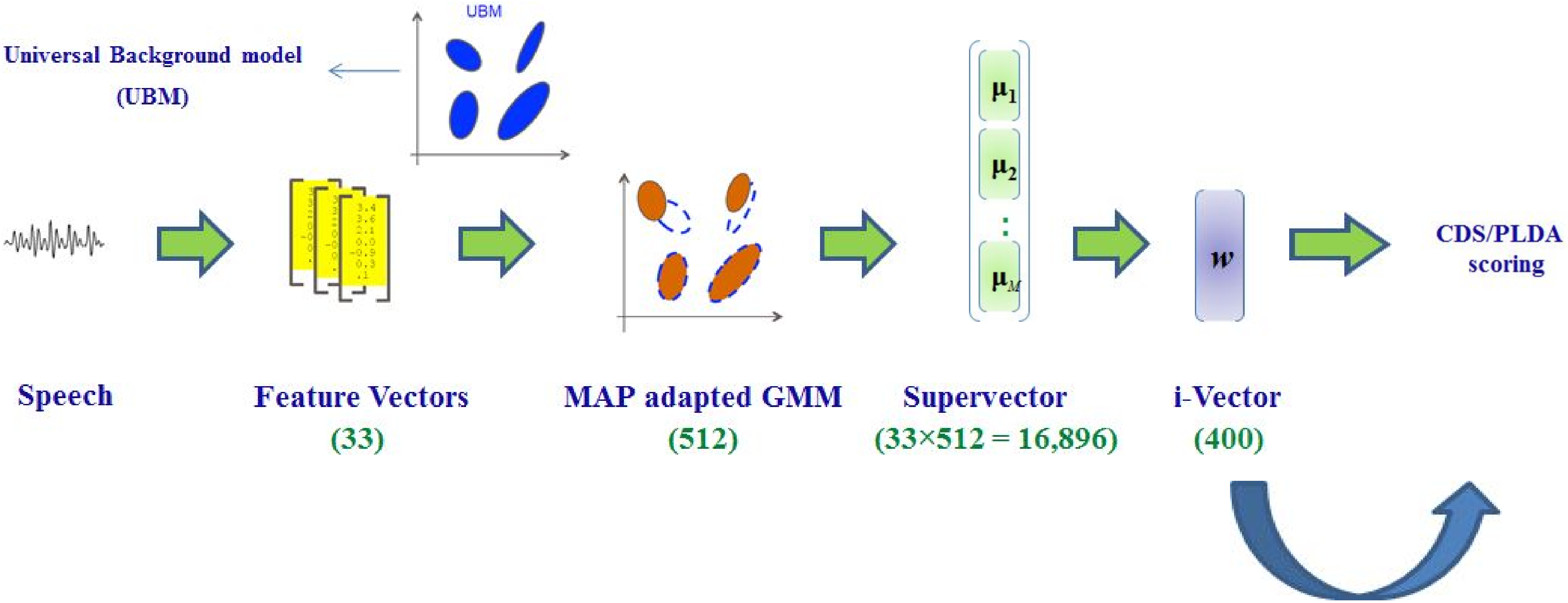
Day 4 Lecture 1

Speaker ID II

Javier Hernando

TALP Research Center

DL Modeling i-Vectors



DL Modeling i-Vectors

Goal :

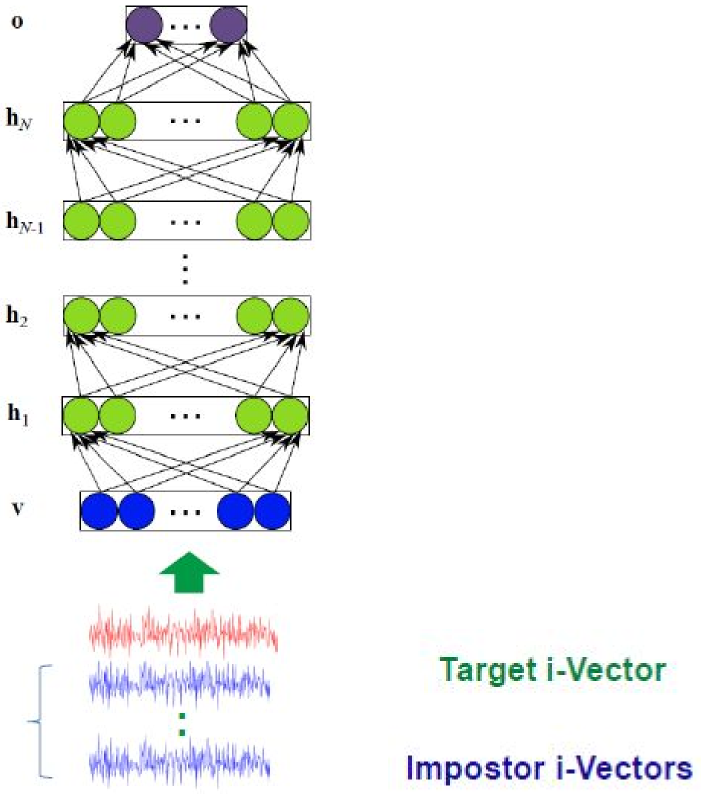
Training a discriminative model for each target speaker

What We Have ?

- **One i-vector (single session) or a couple of i-vectors (multi session) per each target speaker**
- **A large number of background i-vectors (impostors)**

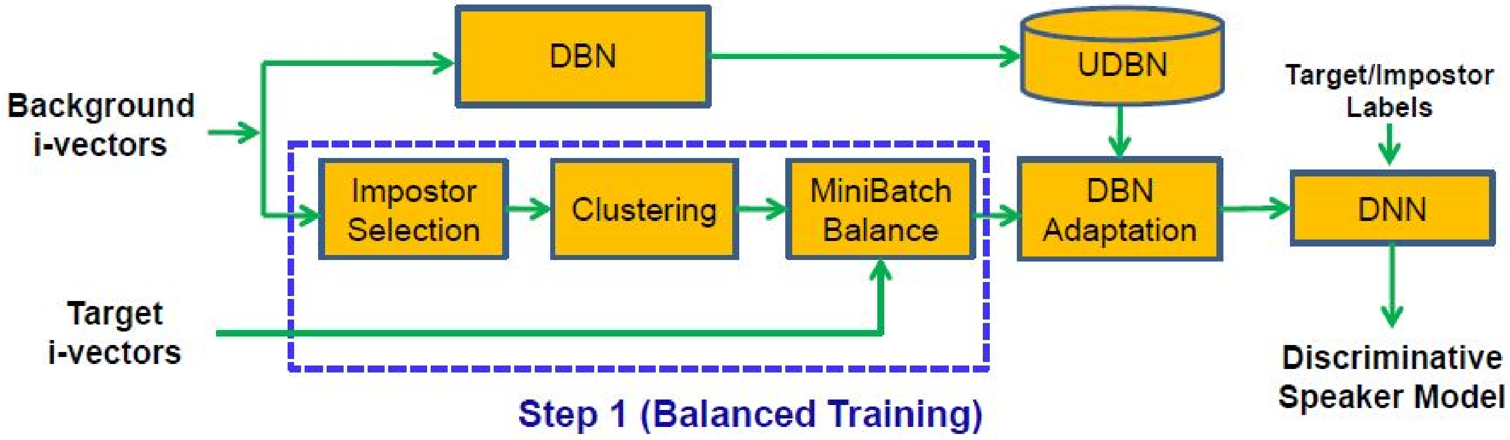
Problems :

- Unbalanced data → Bias towards the majority class
- Few data → Overfitting



O. Ghahabi, J. Hernando, Deep Learning Backend for Single and Multi-Session i-Vector Speaker Recognition, to be appear in IEEE Trans. Audio, Speech and Language Processing

Decoder



Step 1 : Balanced Training

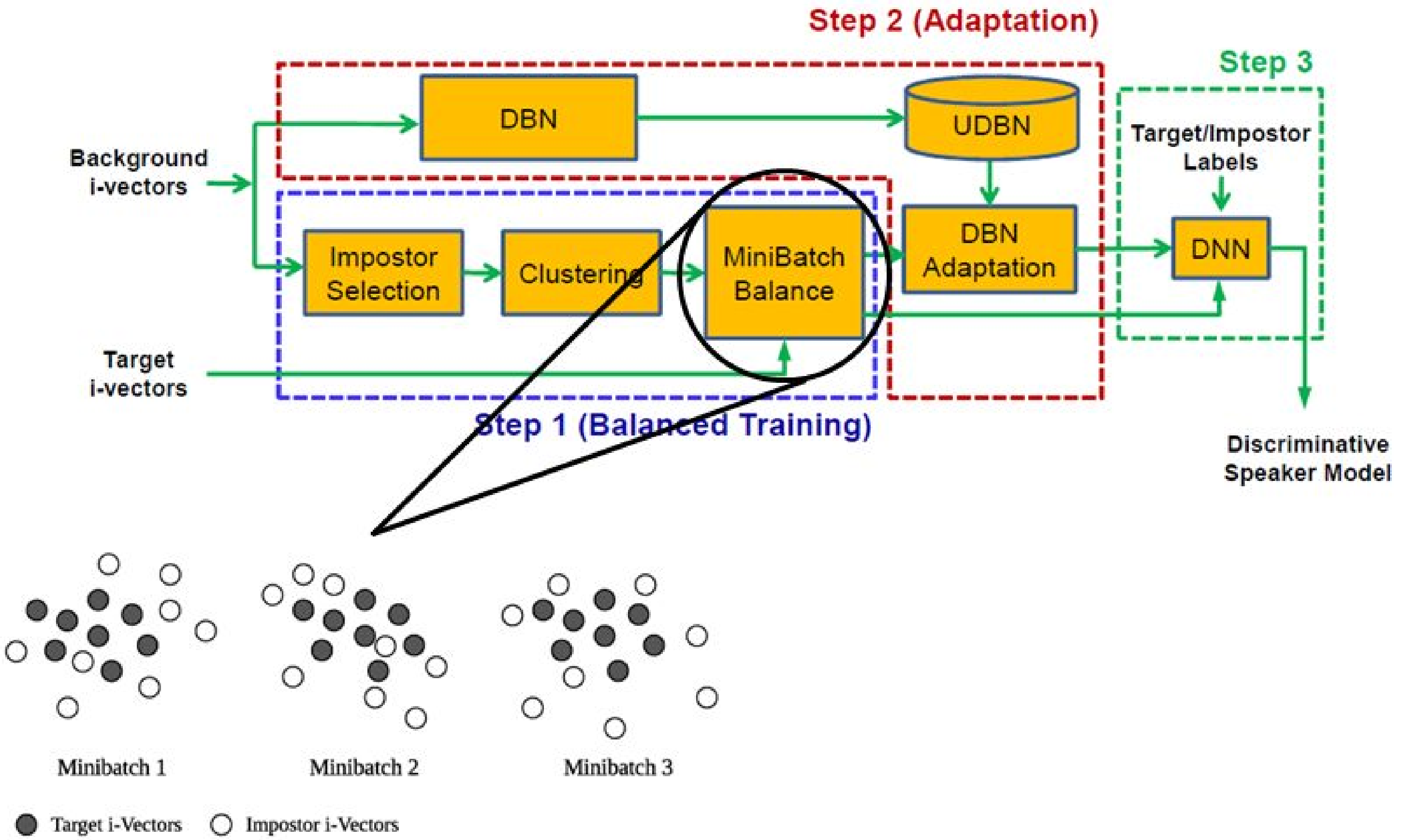
Problem:

A large number of impostor data (negative samples)
Very few number of target data (positive samples)

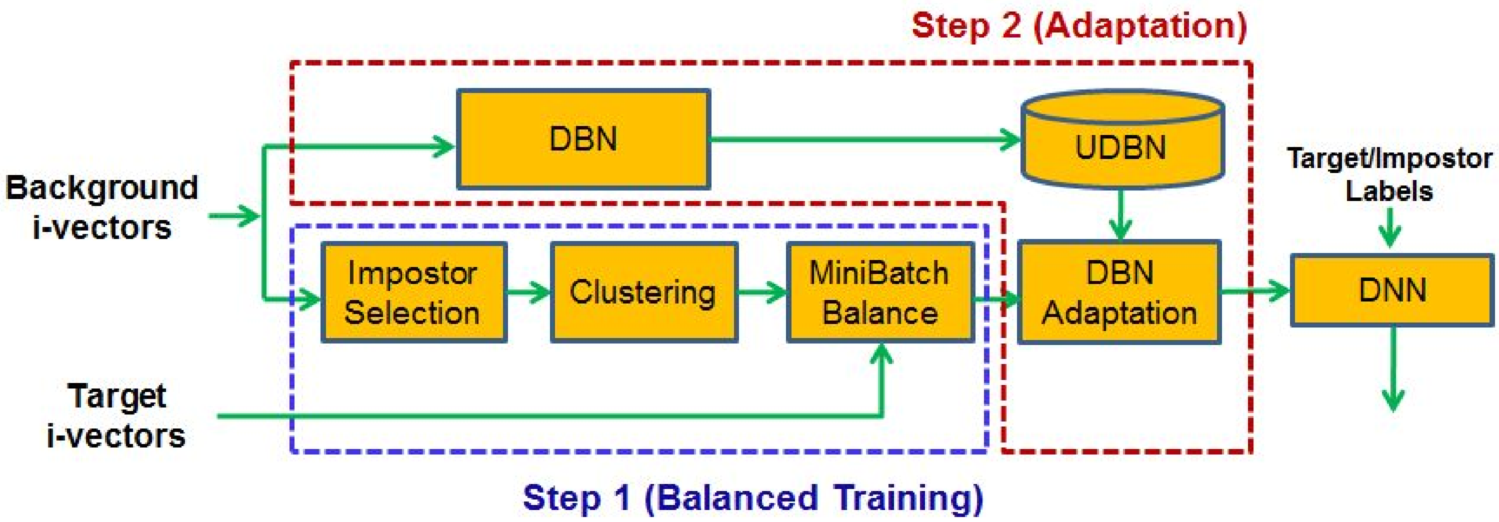
Solutions:

Global Impostor Selection
Clustering using K-means
Equally distributing positive and negative samples among minibatches

DL Modeling i-Vectors



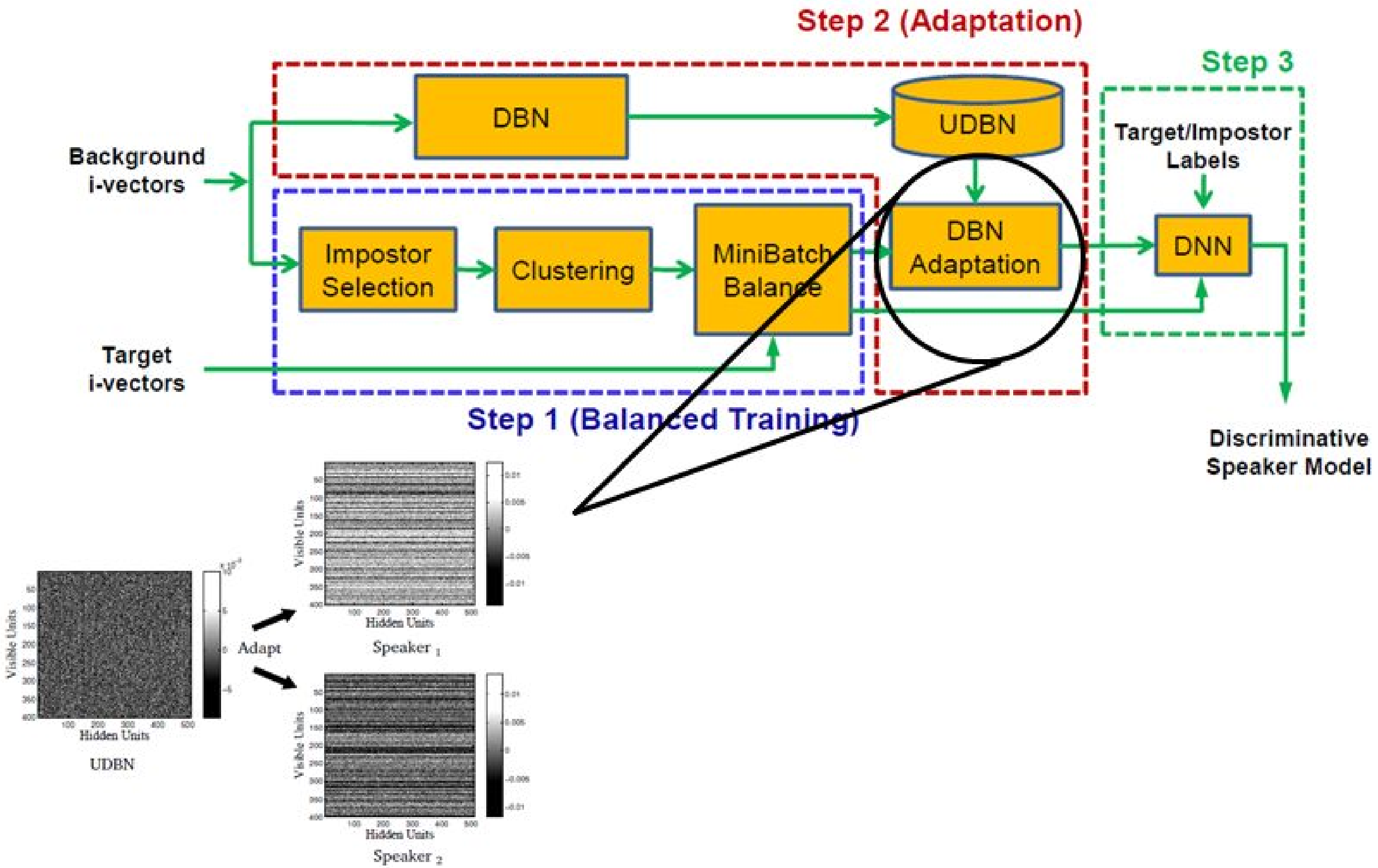
DL Modeling i-Vectors



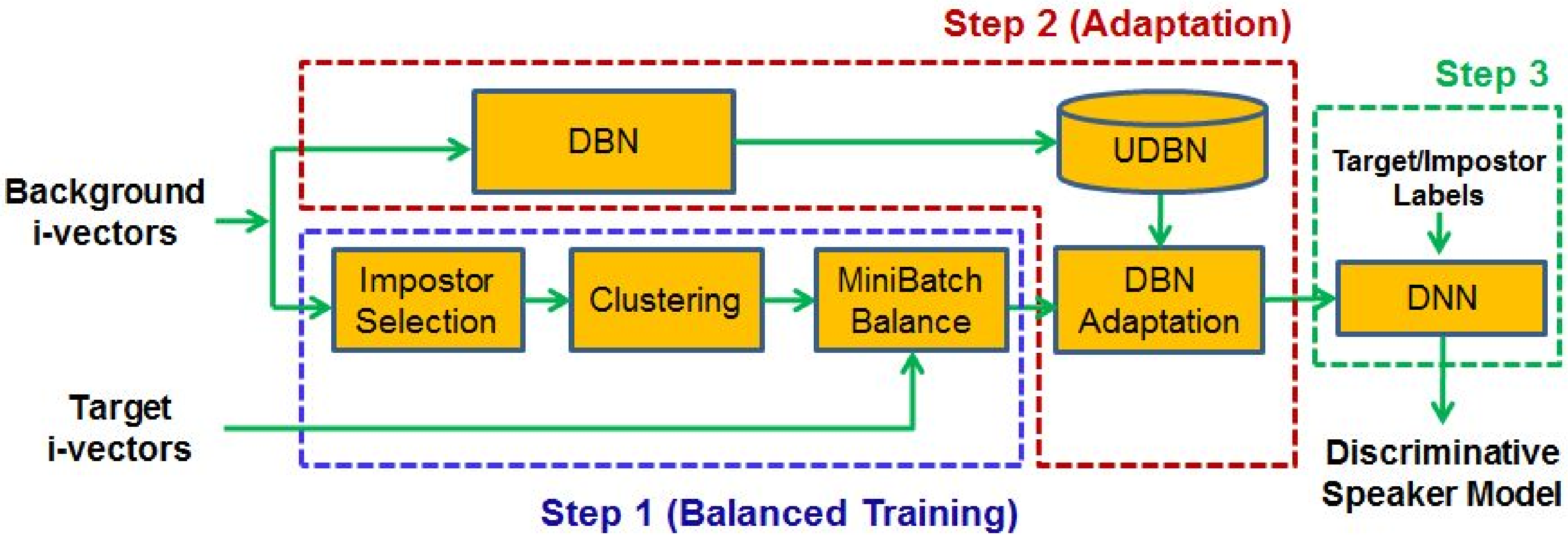
Step 2 : Adaptation

- Universal DBN (Unsupervised learning using background i-vectors)
- Unsupervised Adaptation
  - ✓ Initialize networks by the UDBN parameters
  - ✓ Unsupervised learning using balanced data with few iterations

DL Modeling i-Vectors



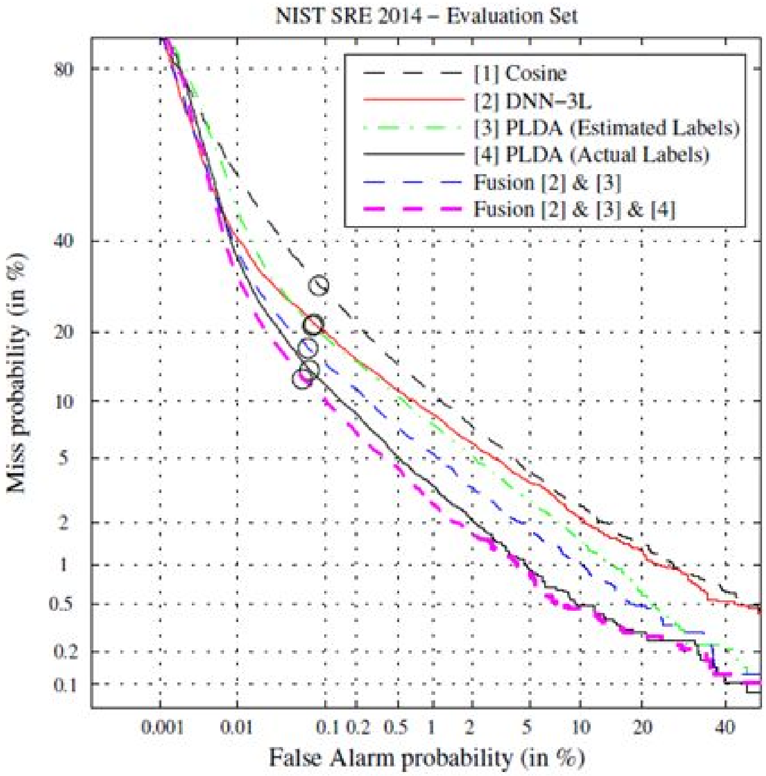
DL Modeling i-Vectors



Step 3 : Fine-Tuning

- Supervised learning given impostor and target labels, adapted DBN, and balanced data

DL Modeling i-Vectors



	Labeled Background Data	Prog Set EER	Prog Set minDCF	Eval Set EER	Eval Set minDCF
[1] Cosine	No	4.78	386	4.46	378
[2] PLDA (Estimated Labels)	No	3.85	300	3.46	284
[3] DNN-3L	No	4.36	297	3.93	291
Fusion [2] & [3]	No	**2.95**	**259**	**2.64**	**238**
[4] PLDA (Actual Labels)	Yes	2.23	226	2.01	207
Fusion [2] & [4]	Yes	2.04	220	1.85	204
Fusion [3] & [4]	Yes	2.10	219	1.98	194
Fusion [2] & [3] & [4]	Yes	**1.90**	**203**	**1.72**	**184**

23% 37%

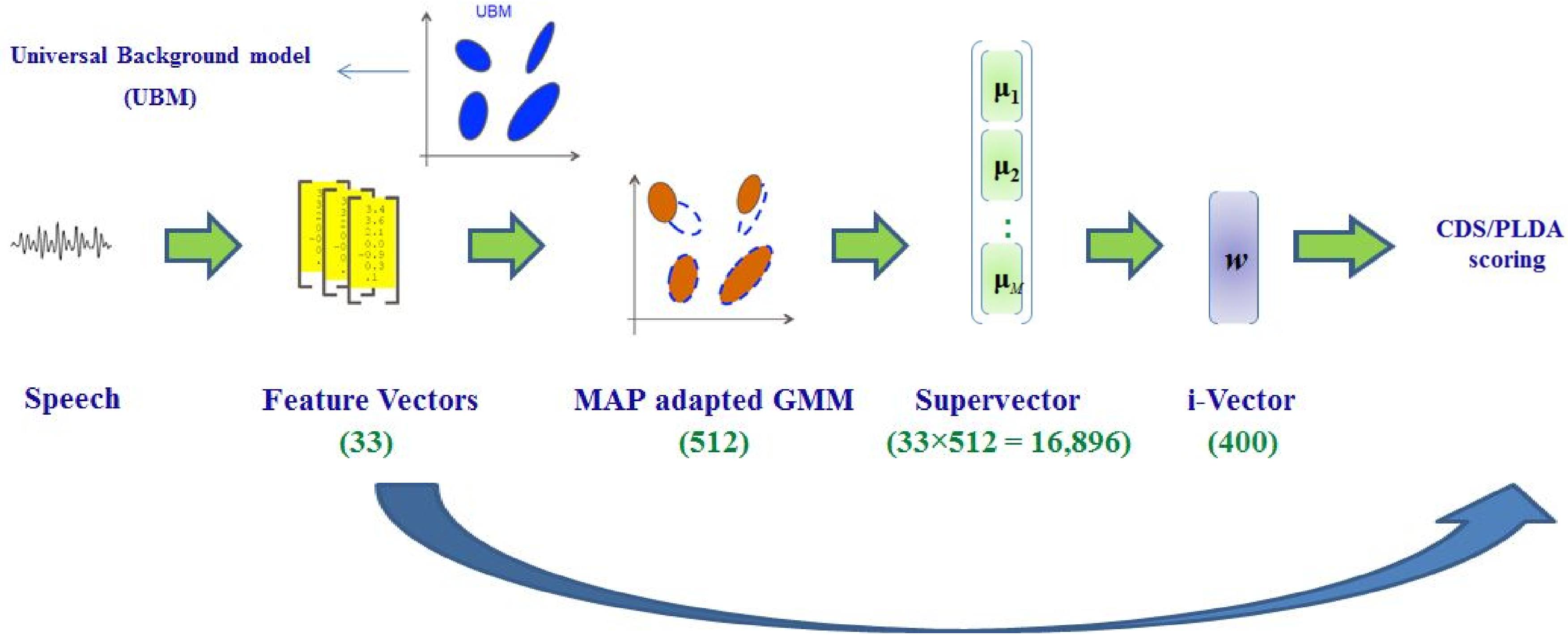
6% 11%

NIST SRE 2014 i-Vector Challenge

(more than 100 participants)

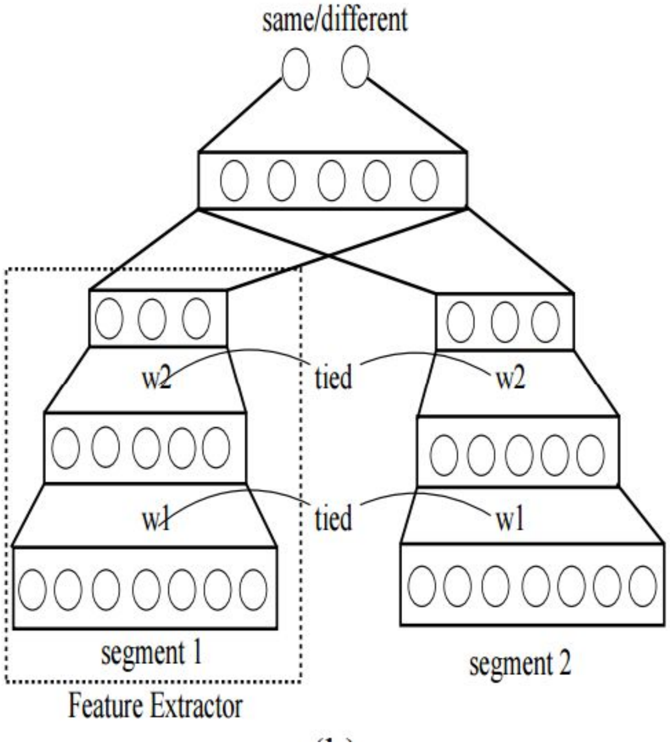
- Top 20 in the 1st Phase (unlabeled background data)
- 2nd rank in the 2nd Phase (labeled background data)

DL Feature Classification

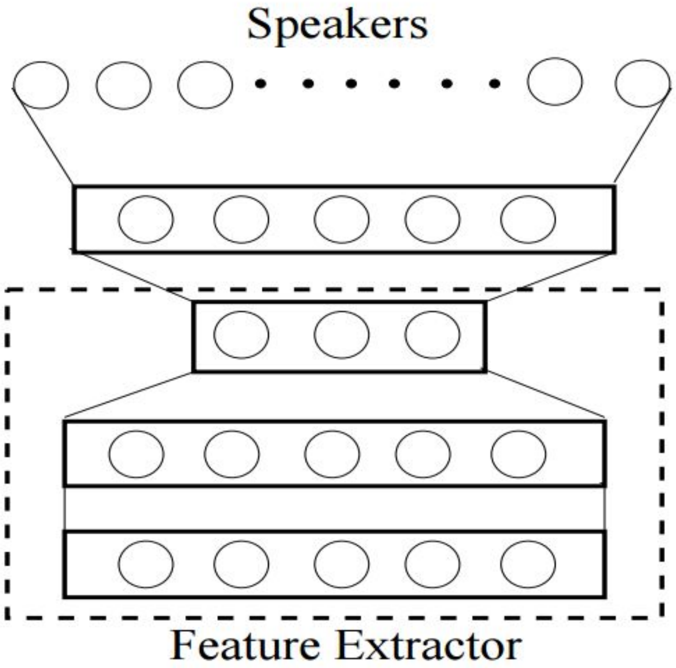


DL Feature Classification

Credit S. H. Yella,

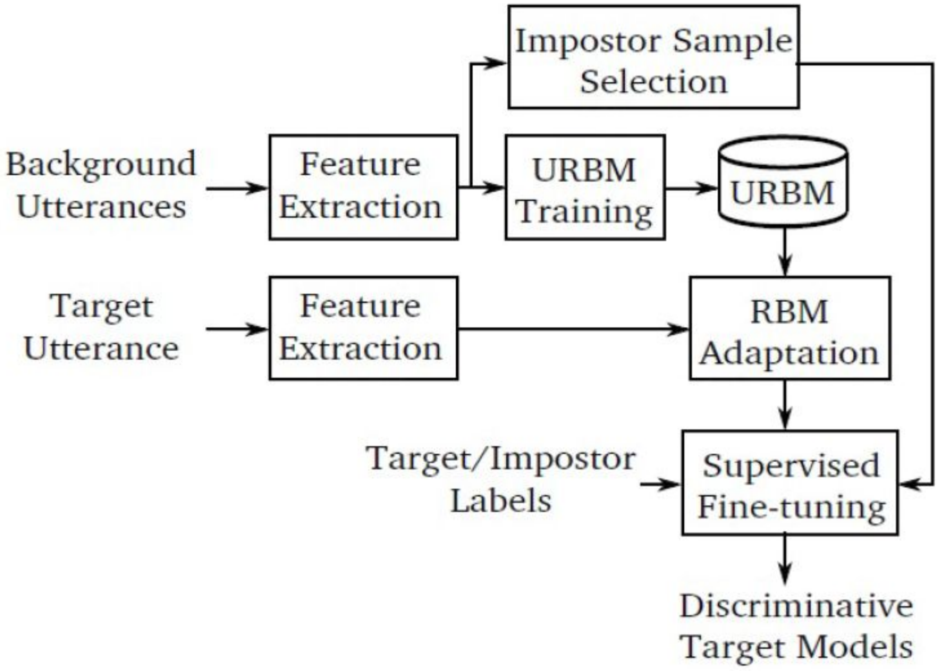


Speaker Verification



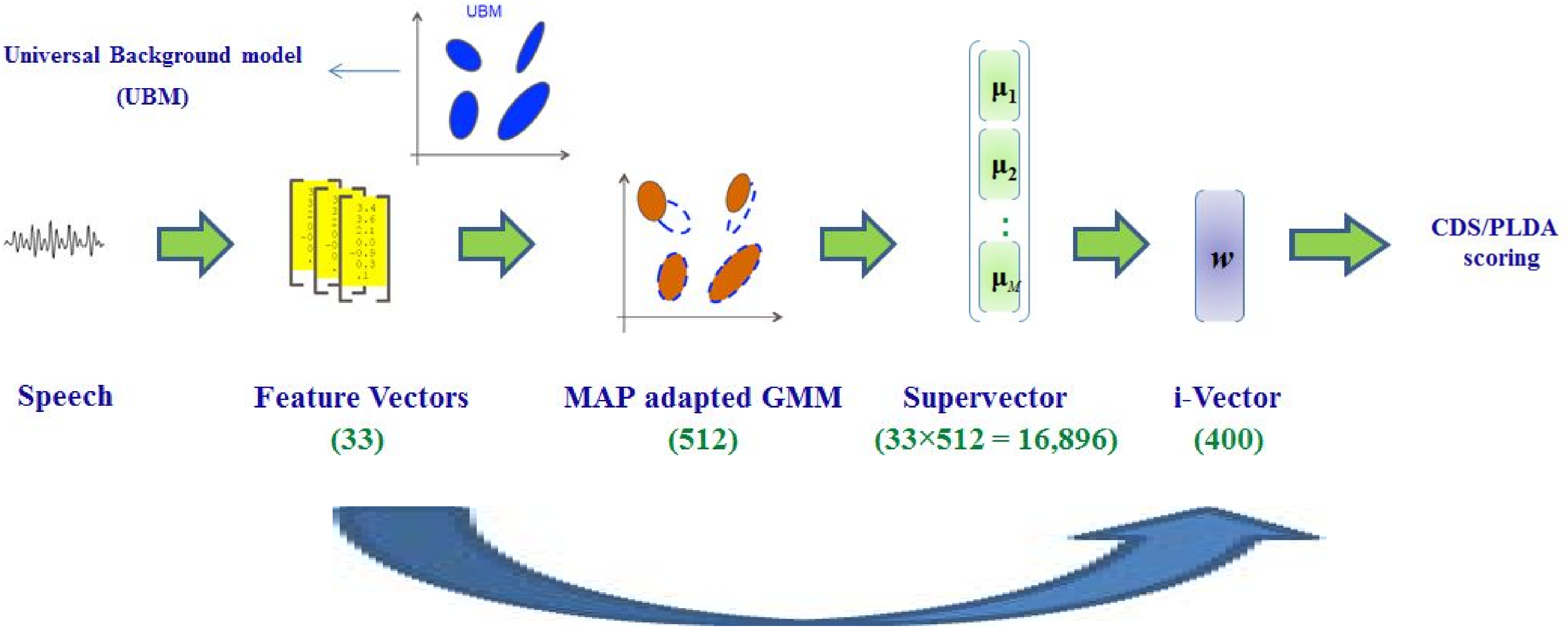
Speaker Identification

DL Feature Classification

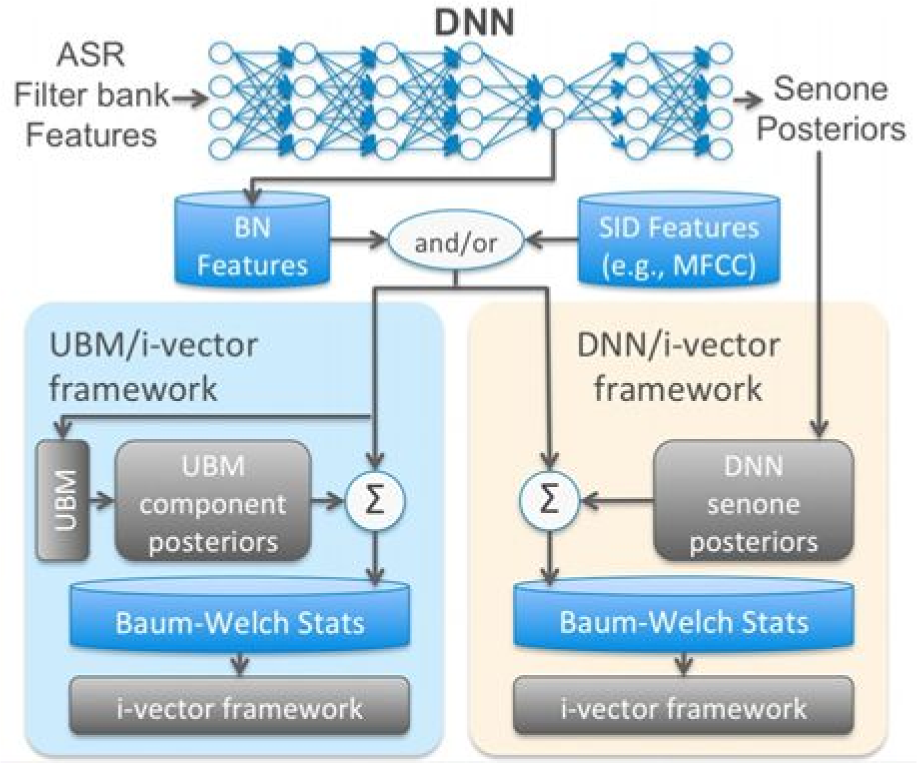


P. Safari, O. Ghahabi, J. Hernando, “Restricted Boltzmann Machines for speaker vector extraction and feature classification”, Proc. URSI 2016

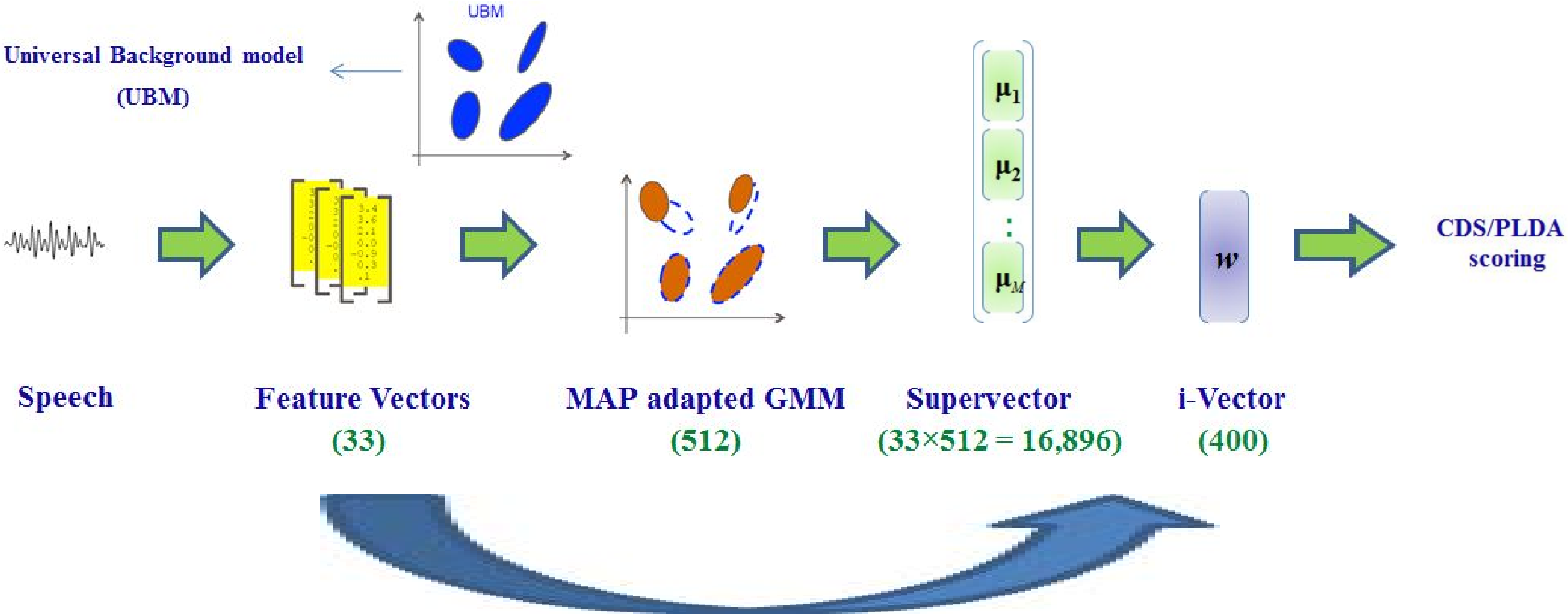
DL i-vector Extraction



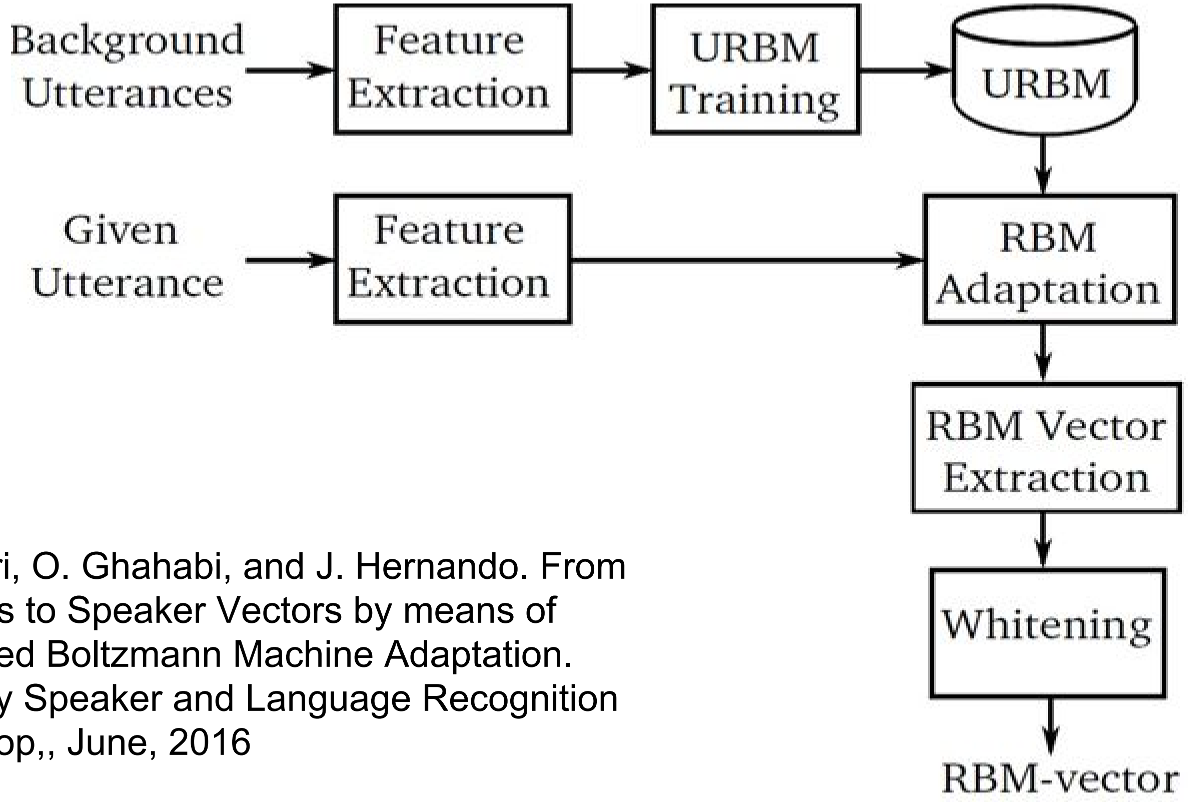
DL i-vector Extraction



DL ‘speaker-vectors’



RBM vectors



P. Safari, O. Ghahabi, and J. Hernando. From Features to Speaker Vectors by means of Restricted Boltzmann Machine Adaptation. Odyssey Speaker and Language Recognition Workshop,, June, 2016

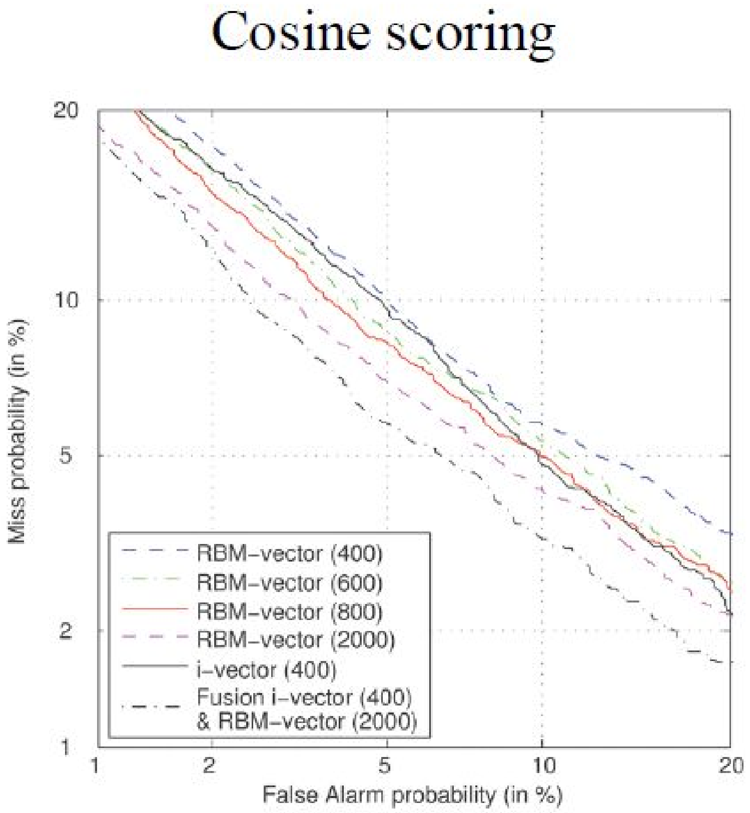
RBM vectors

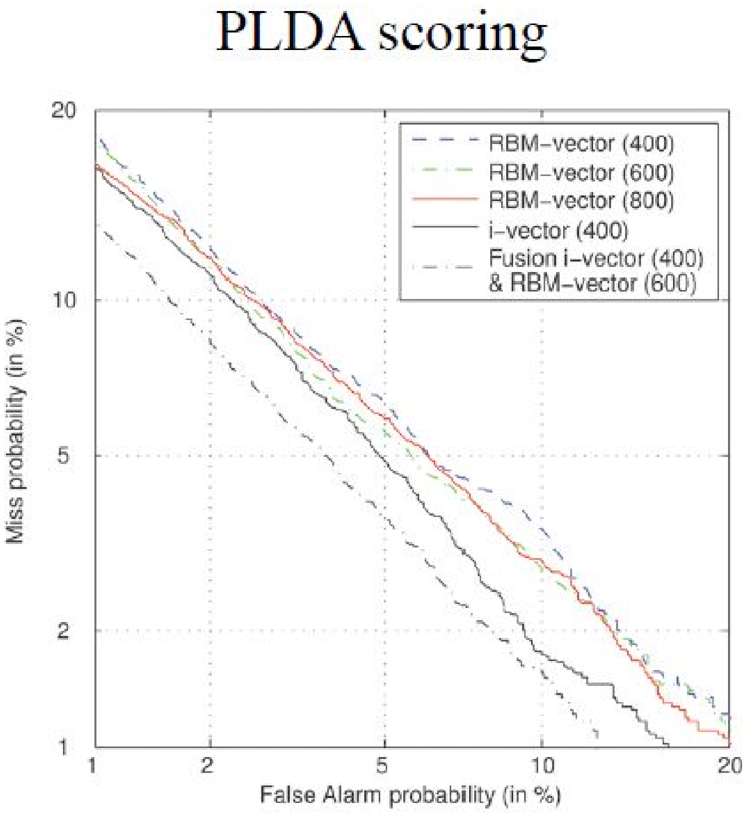
- RBM supervectors

$$\begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1h} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{v1} & \omega_{v2} & \cdots & \omega_{vh} \end{pmatrix} \qquad \begin{pmatrix} \begin{pmatrix} \omega_{11} \\ \omega_{12} \\ \vdots \\ \omega_{1h} \end{pmatrix} \\ \begin{pmatrix} \omega_{21} \\ \omega_{22} \\ \vdots \\ \omega_{2h} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \omega_{v1} \\ \omega_{v2} \\ \vdots \\ \omega_{vh} \end{pmatrix} \end{pmatrix}$$

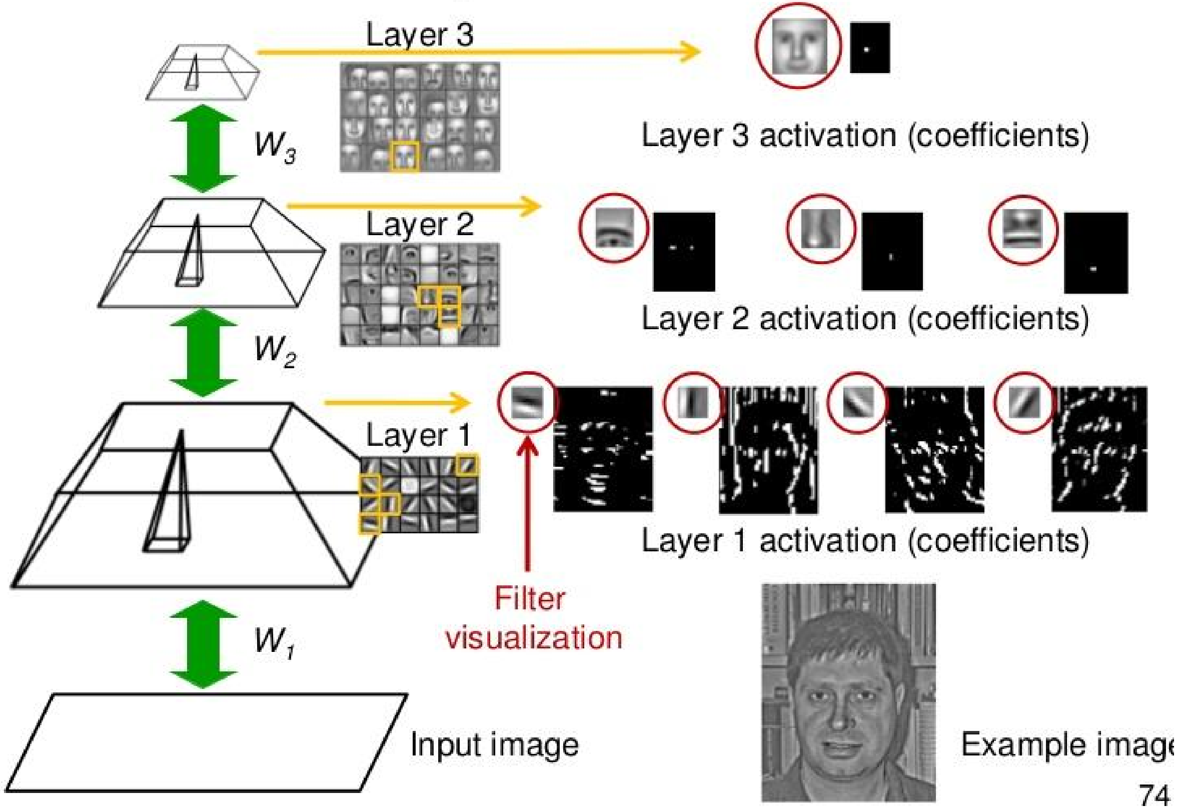
- Mean-normalization
- PCA whitening with dimension reduction
- PCA trained based on all background RBM supervectors
- The output of the whitening stage is called RBM-vector

RBM vectors

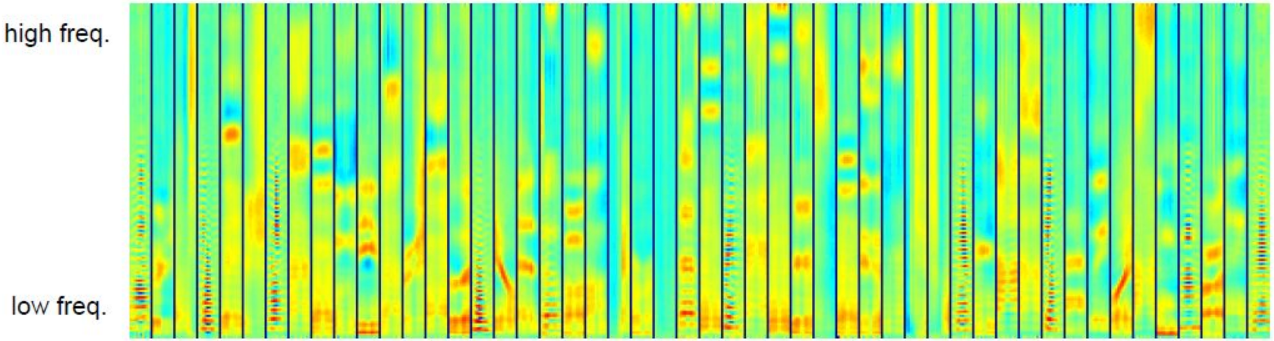

Cosine scoring
20
10
5
2
1
Miss probability (in %)
1
2
5
10
20
False Alarm probability (in %)
RBM–vector (400)
RBM–vector (600)
RBM–vector (800)
RBM–vector (2000)
i–vector (400)
Fusion i–vector (400)
& RBM–vector (2000)


PLDA scoring
20
10
5
2
1
Miss probability (in %)
1
2
5
10
20
False Alarm probability (in %)
RBM–vector (400)
RBM–vector (600)
RBM–vector (800)
i–vector (400)
Fusion i–vector (400)
& RBM–vector (600)

CDBN vectors



CDBN vectors



randomly selected first-layer CDBN bases

Unsupervised feature learning for audio classification using convolutional deep belifef networks, H. Lee et al., Advances in Neural Information Processing Systems, 22:1096–1104, 2009

DL 'supervector like' estimation

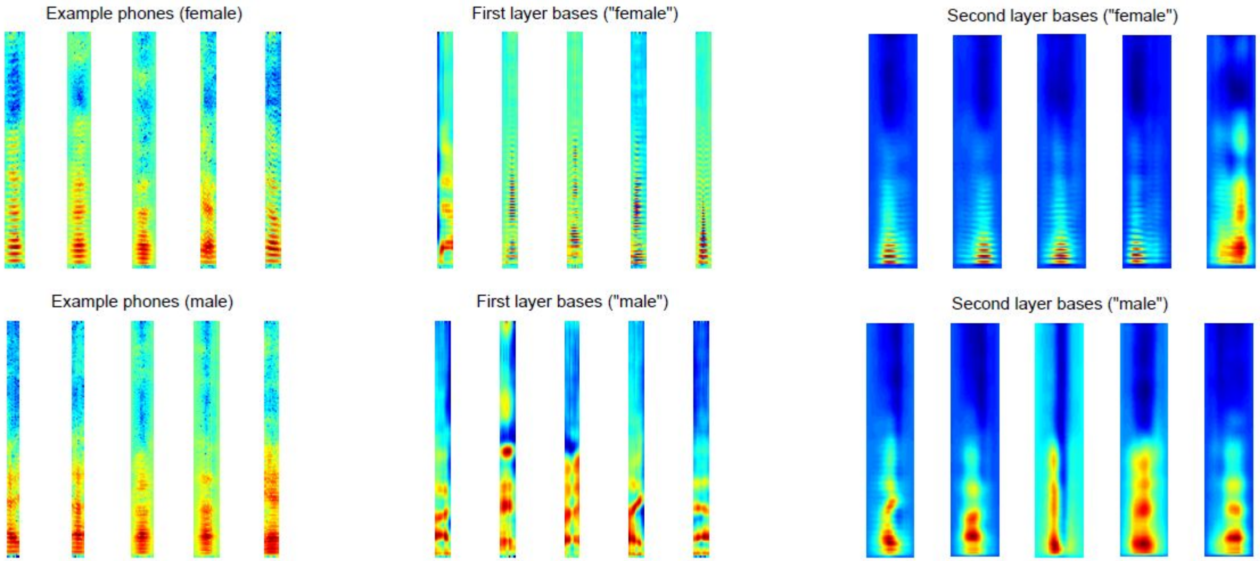
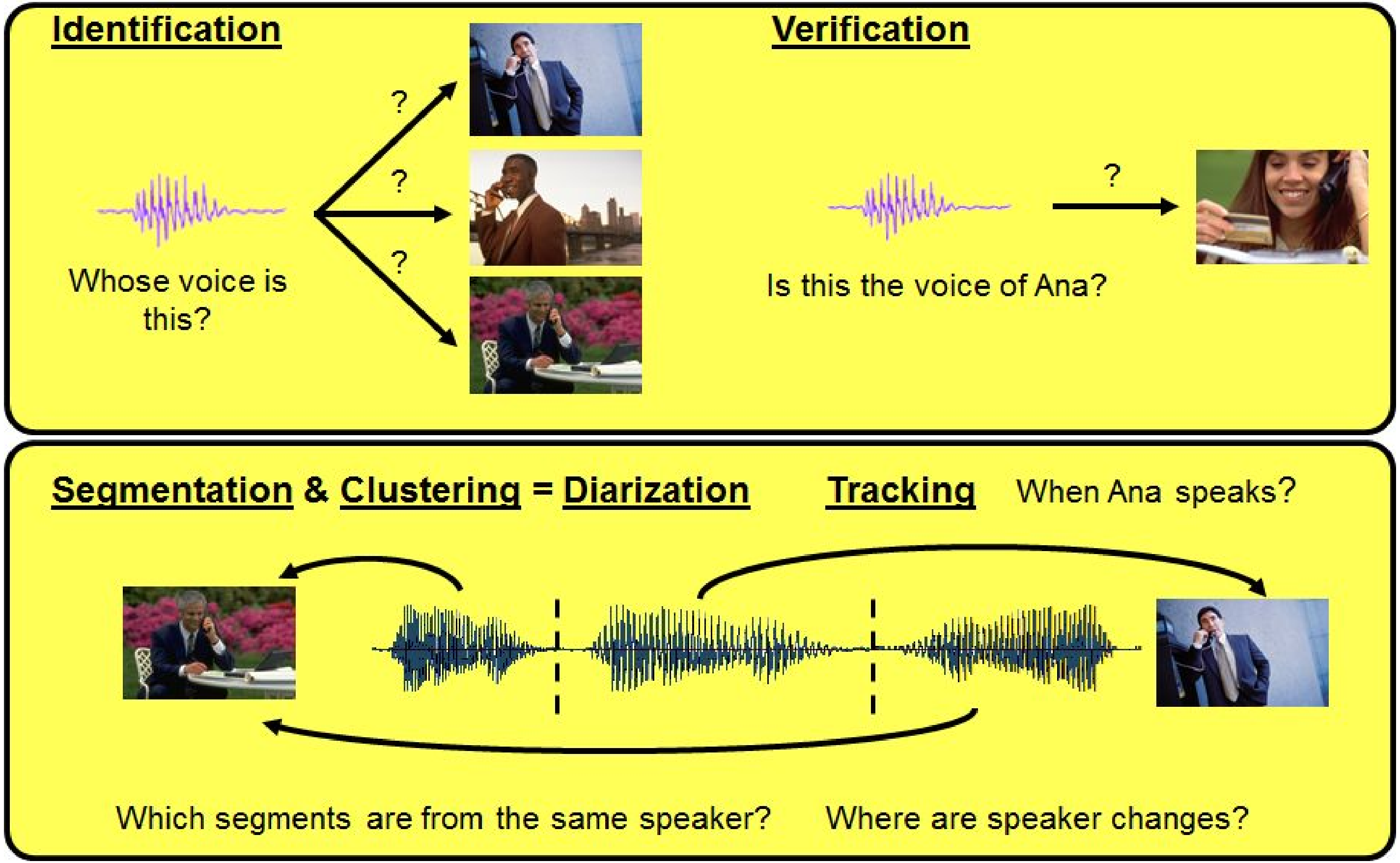
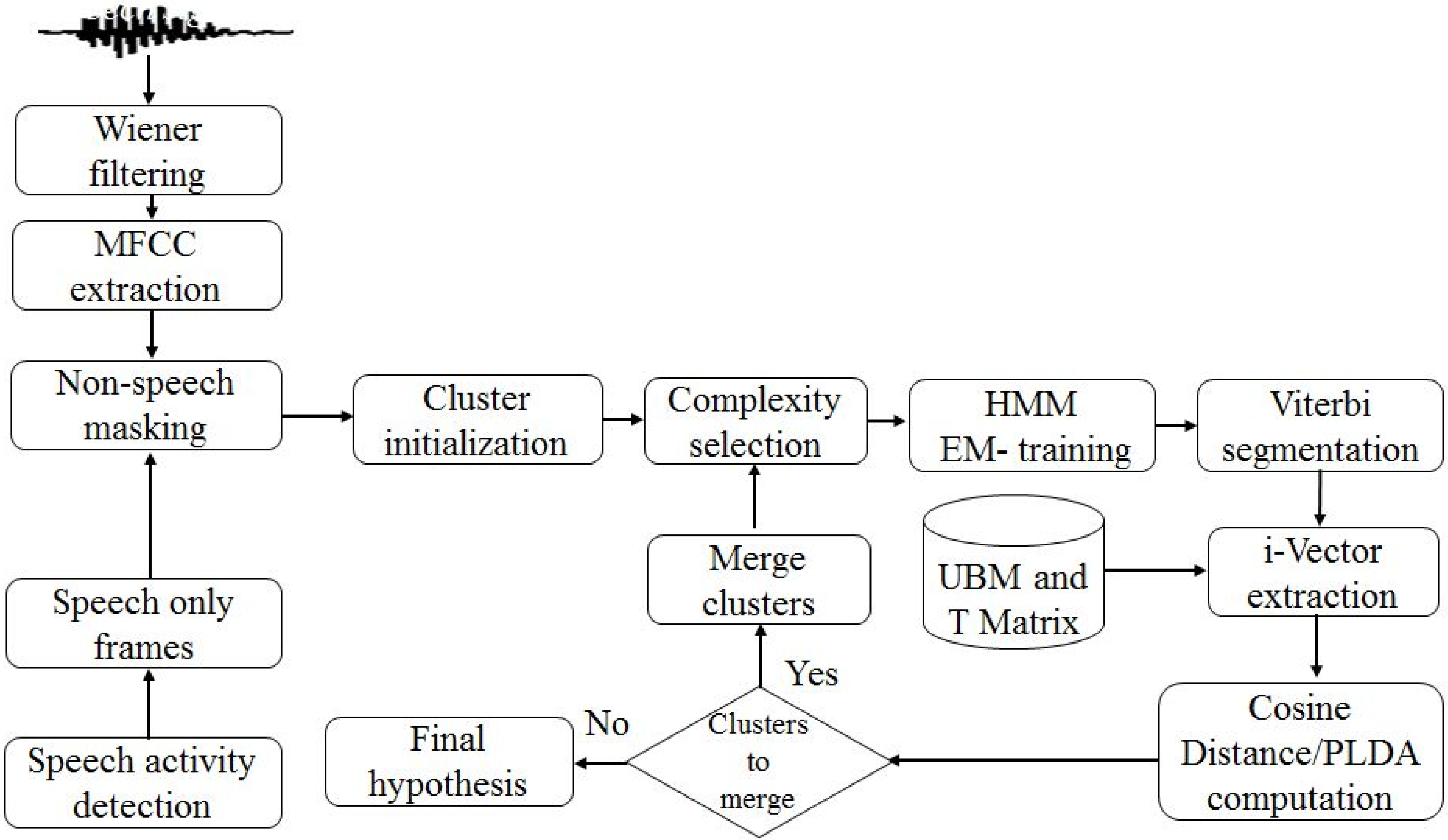


Figure 3: (Left) five spectrogram samples of "ae" phoneme from female (top)/male (bottom) speakers. (Middle) Visualization of the five first-layer bases that most differentially activate for female/male speakers. (Right) Visualization of the five second-layer bases that most differentially activate for female/male speakers.

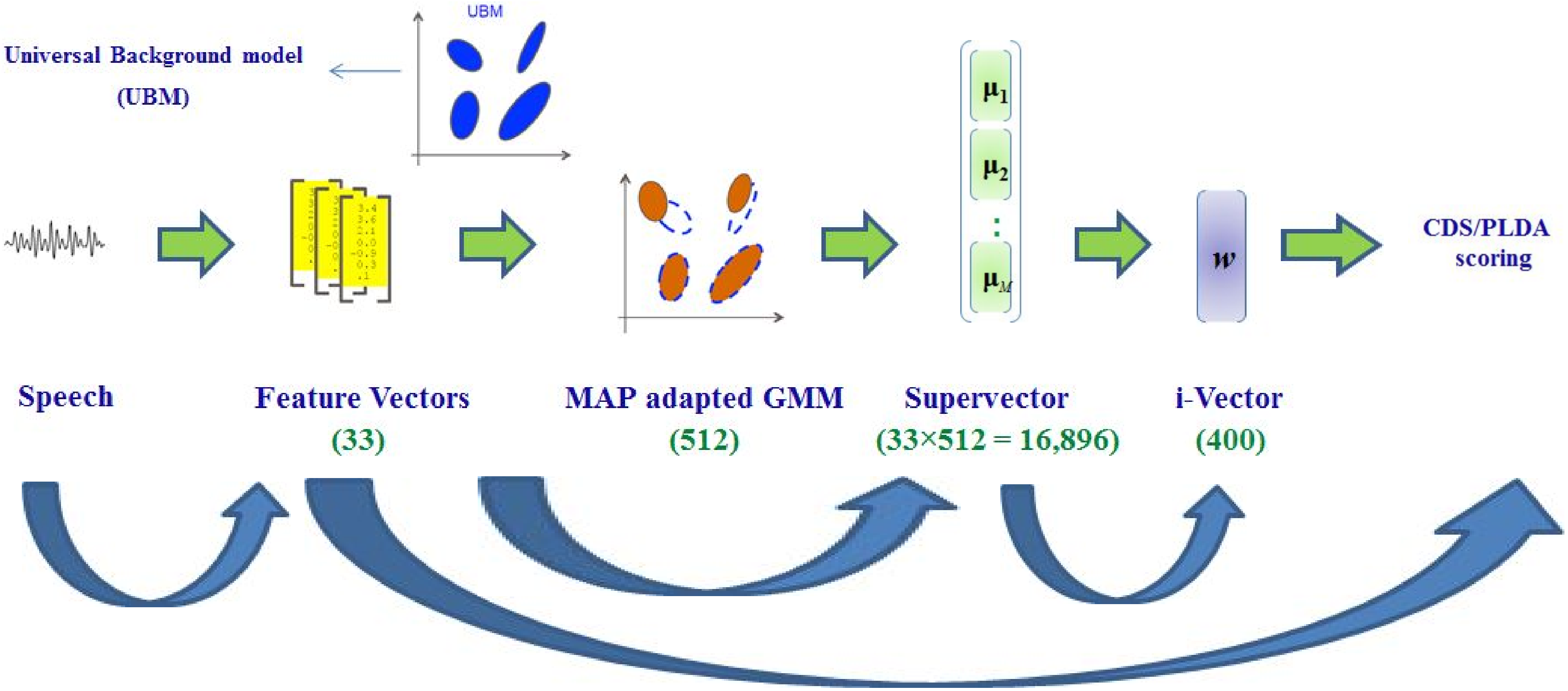
Tasks


Identification
?
?
?
Whose voice is this?
Verification
?
Is this the voice of Ana?
Segmentation & Clustering = Diarization
Tracking
When Ana speaks?
Which segments are from the same speaker?
Where are speaker changes?

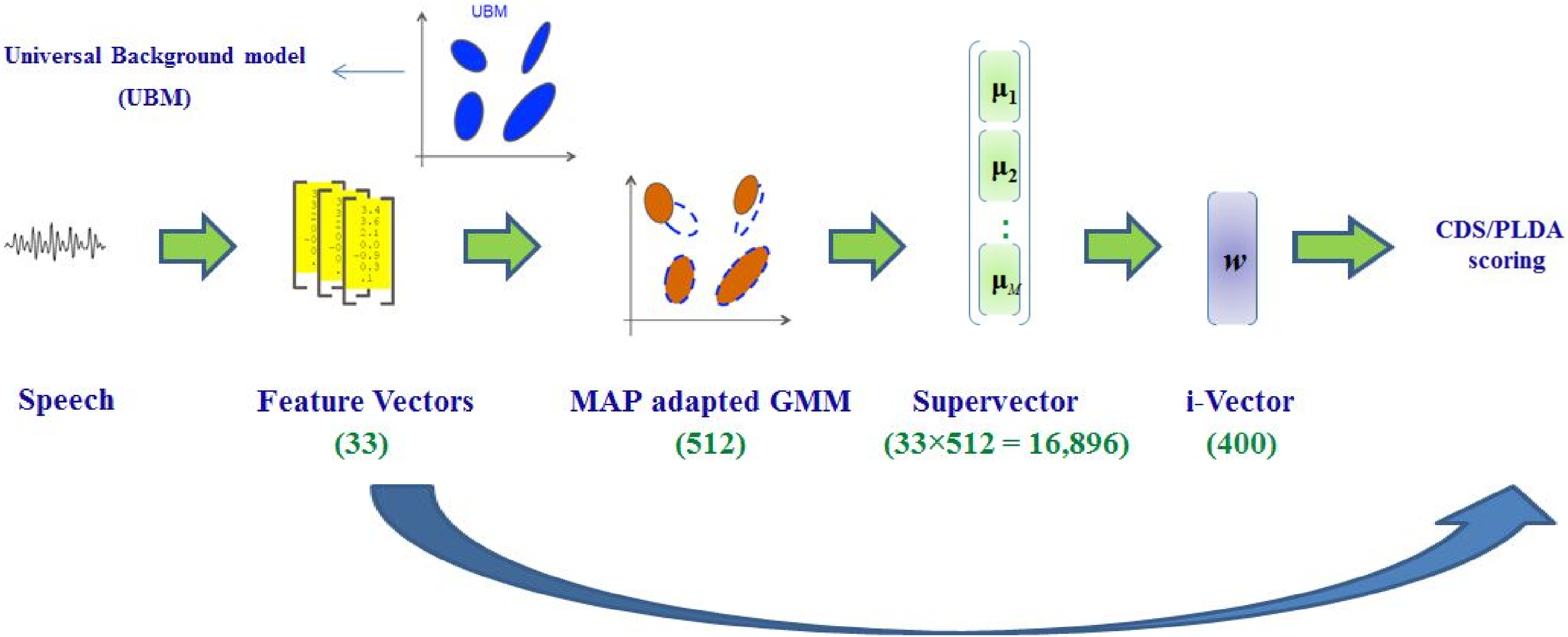
SoA Speaker Diarization



DL in Speaker Diarization



DL Feature Classification



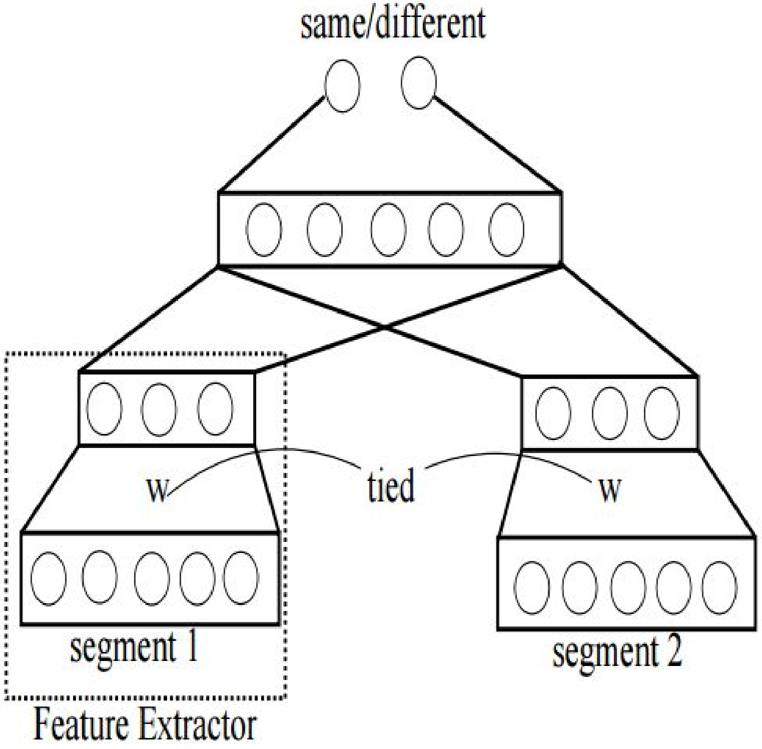
Speaker Clustering: Speaker Comparison

Pa

Harsha et al. "Artificial Neural Network Features for Speaker Diarization". IEEE Spoken Language Technology Workshop. (2014) 402-406

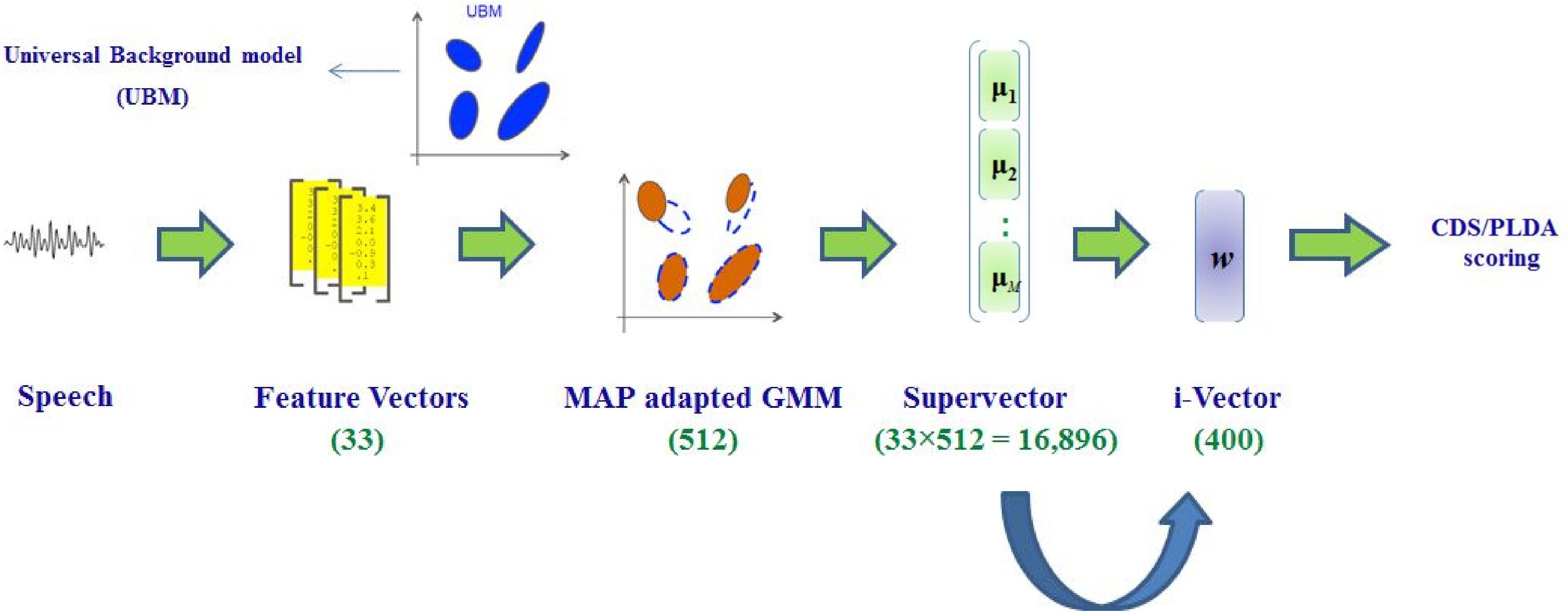
Train	Test	MFCC	ANN + MFCC	Rel. change
AMI	AMI	25.1	21.5	-14.3%
AMI	ICSI	20.6	18.4	-10.7%
ICSI	ICSI	20.6	15.1	-26.7%

Speaker errors obtained on AMI and ICSI datasets for matched and mismatched training conditions. MFCC corresponds to baseline clustering using BIC. ANN+MFCC is referred to the ANN shown in right figure.

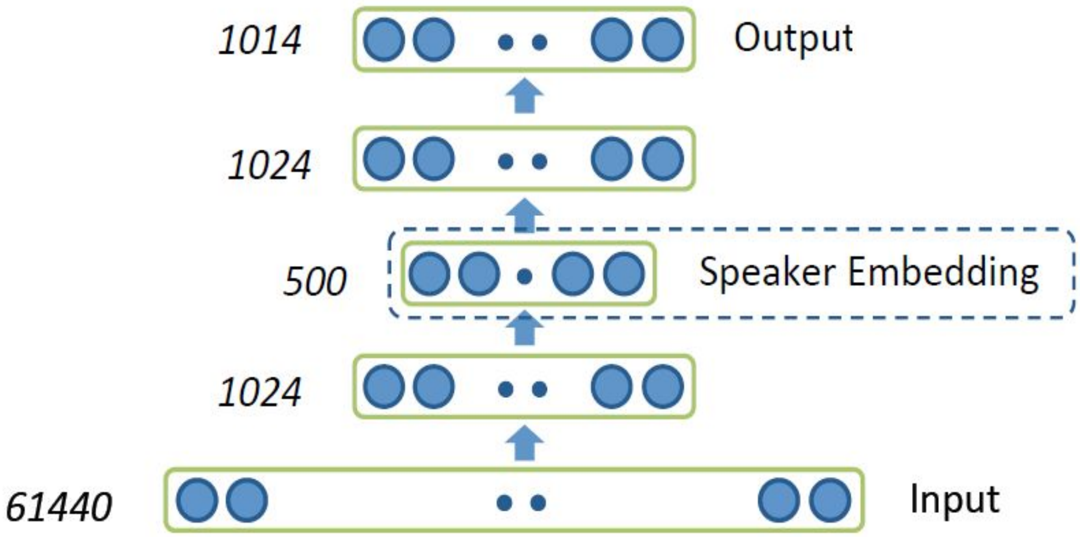


Shallow Speaker Comparison

DL 'speaker-vectors'



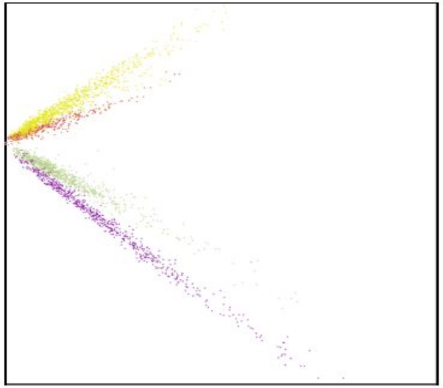
Speaker Embeddings



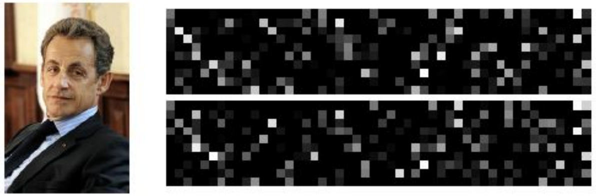
$$s_g = \frac{1}{\sum_t \gamma_g(t)} \sum_t \gamma_g(t)(x_t - \mu_g)$$

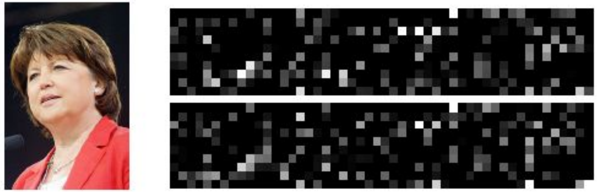
Mickael Rouvier et al. "Speaker Diarization trough Speaker Embeddings". 23rd European Signal Processing Conference. (2015)

Speaker Embeddings



2D projection of four Speaker Embeddings using PCA.



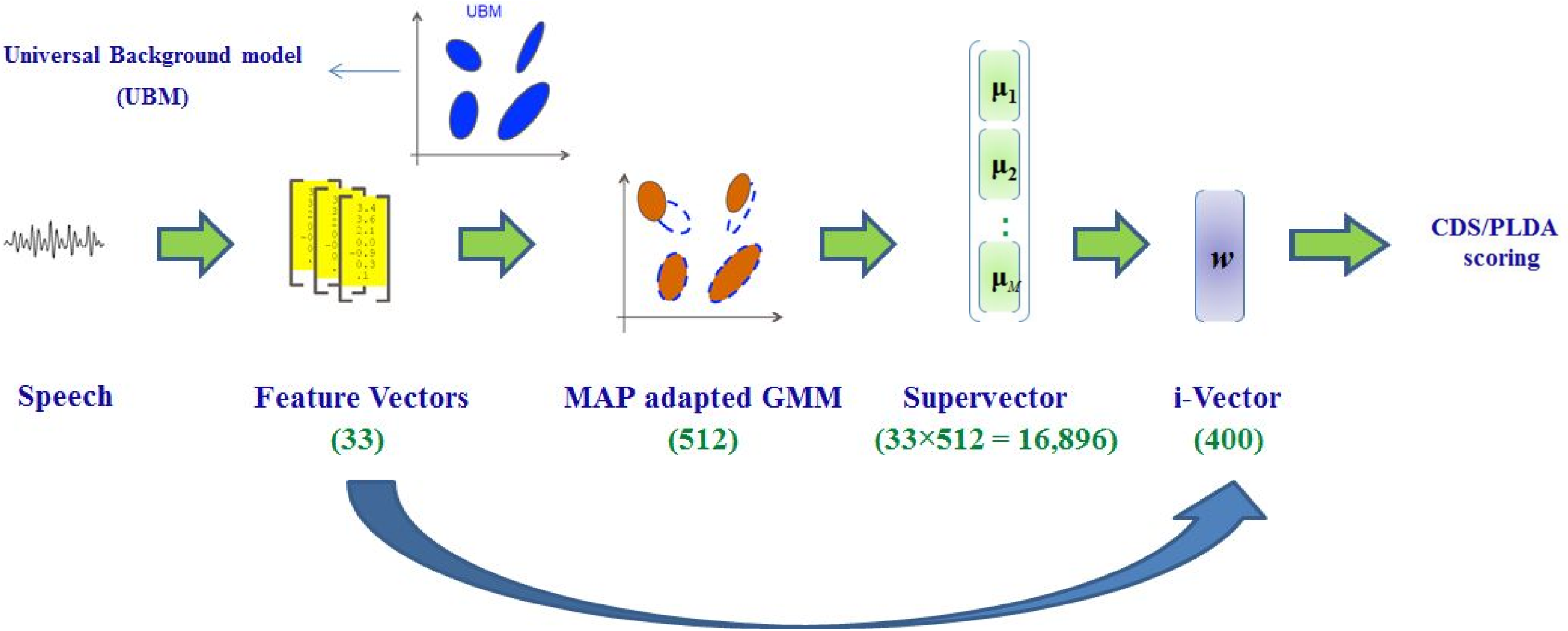


500 size Speaker Embeddings rearranged in 10x50. Representation of two utterances from each speaker.

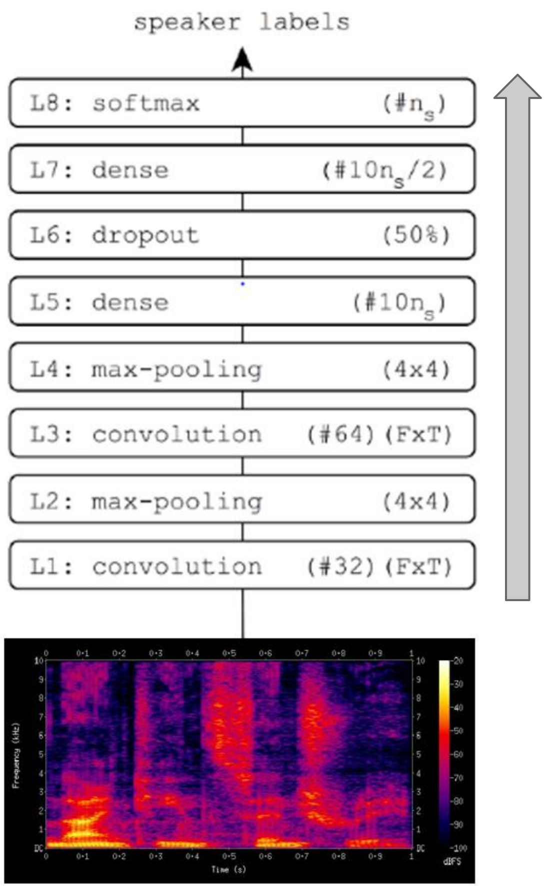
Layer/Dim	300	400	500	600	700
Layer_1	22.11	22.38	20.80	20.10	21.78
Layer_2	21.26	21.08	**20.15**	20.52	20.79
Layer_3	23.97	19.58	21.44	21.73	21.78

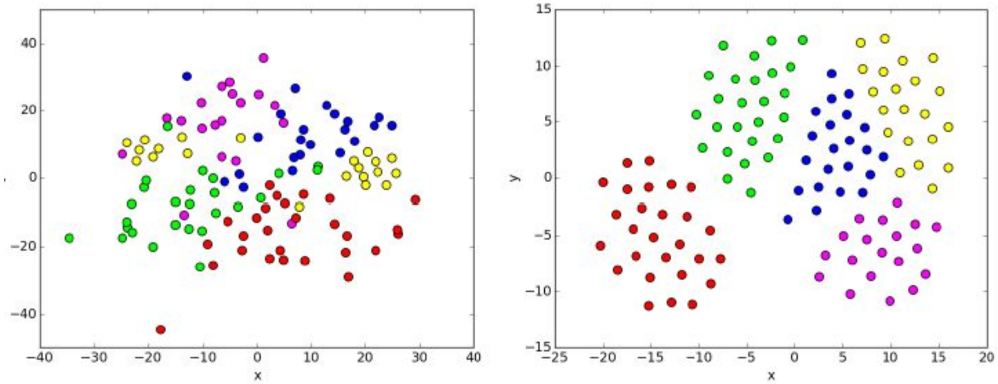
$$DER = \frac{\#Spk + \#Miss + \#FA}{\#Total}$$

DL ‘speaker-vectors’



CNN BN Feature





Five Speaker representations in 2 dimensions.
Left figure show the output vector of the softmax layer L8.
Right figure correspond to the same output vector of L5 dense layer.
Differents colors are assigned to different speakers.

Yanik Lukic et al. "Speaker Identification and Clustering using Convolutional Neural Networks". In 2016 IEEE International workshop on machine learning for signal processing. (2016)

CNN BN Features

- L5 and L7 size depend proportionally to the number of speakers.
- L5 and L7 outperforms the softmax layer L8, where L7 is better than L5.
- trainning data (speaker ammount) must be above 10 * (# speakers) for a good performance.

	20 speakers		**40 speakers**	
Layer	**MR 100**	**MR 590**	**MR 100**	**MR 590**
L5: dense	0.100	0.100	0.300	0.125
L7: dense	0.100	**0.100**	0.325	**0.050**
L8: softmax	0.450	0.250	0.700	0.450

$$MR = \frac{1}{N}\sum_{j=1}^{N_s} e_j.$$