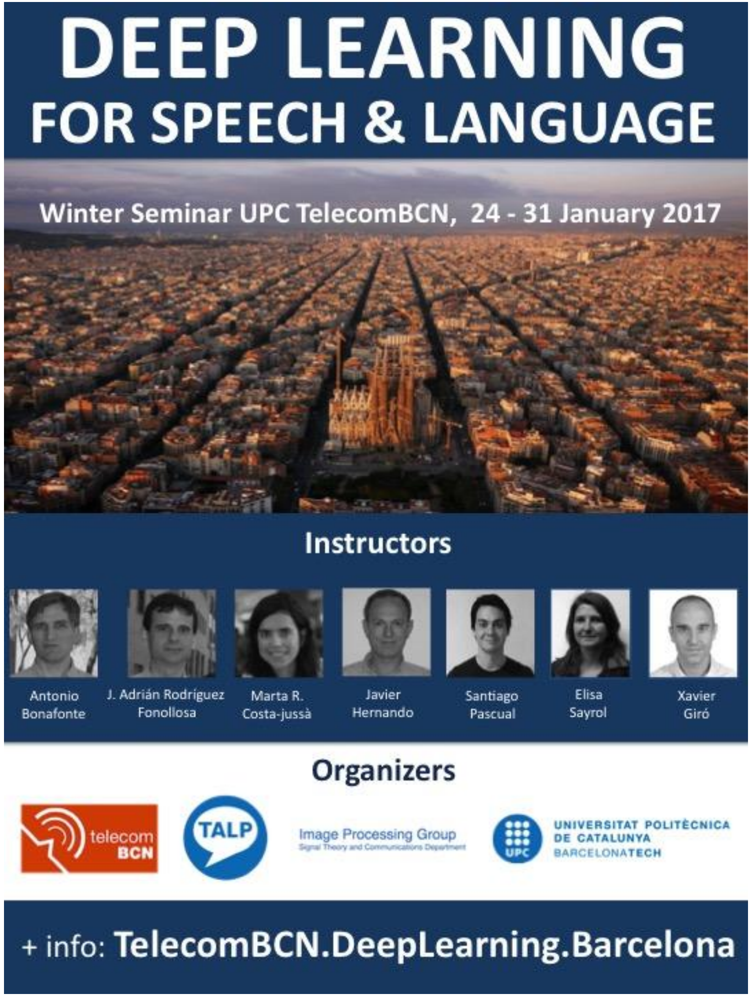


[course site]

Day 3 Lecture 3

Speaker ID I

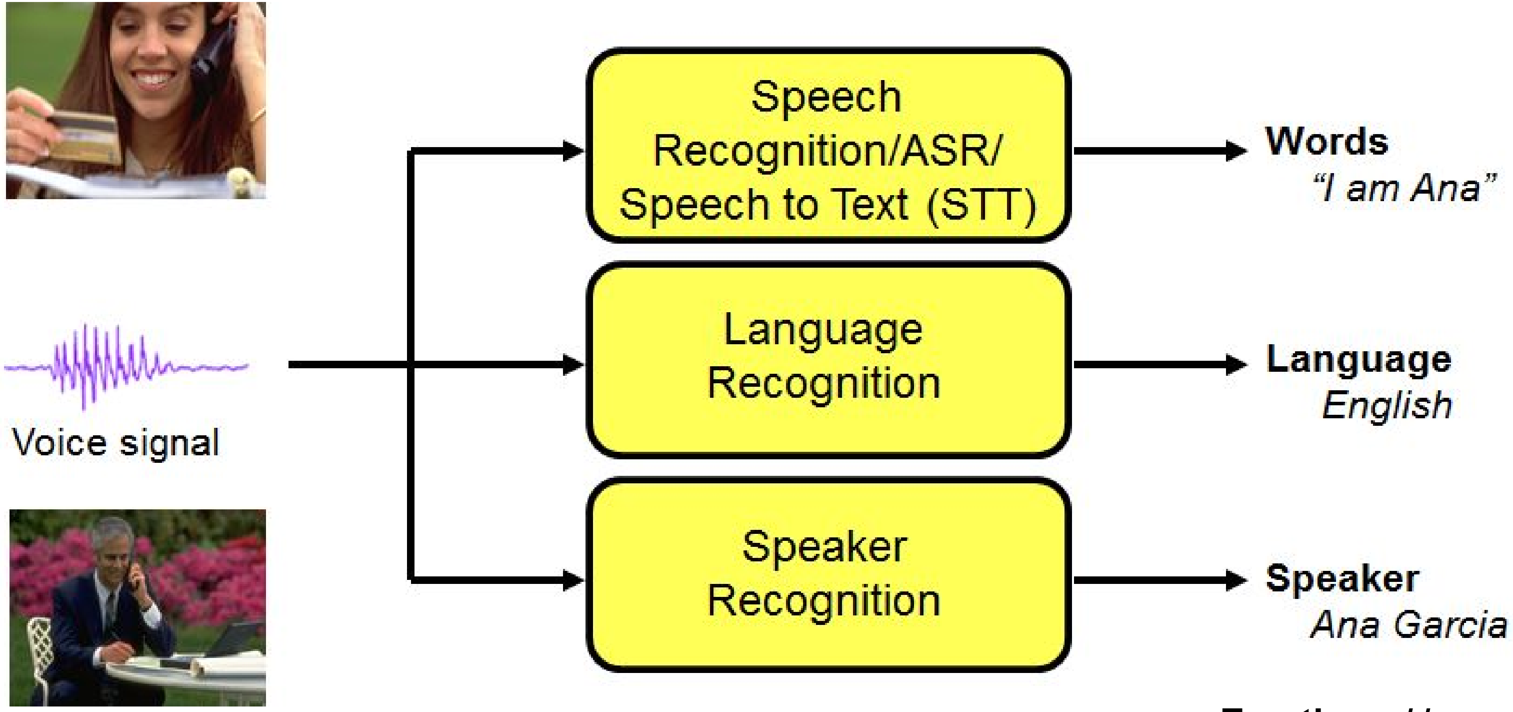
Javier Hernando

TALP Research Center

Acknowledgments

Miquel India, Omid Ghahabi, Pooyan Safari
Ph.D. candidates

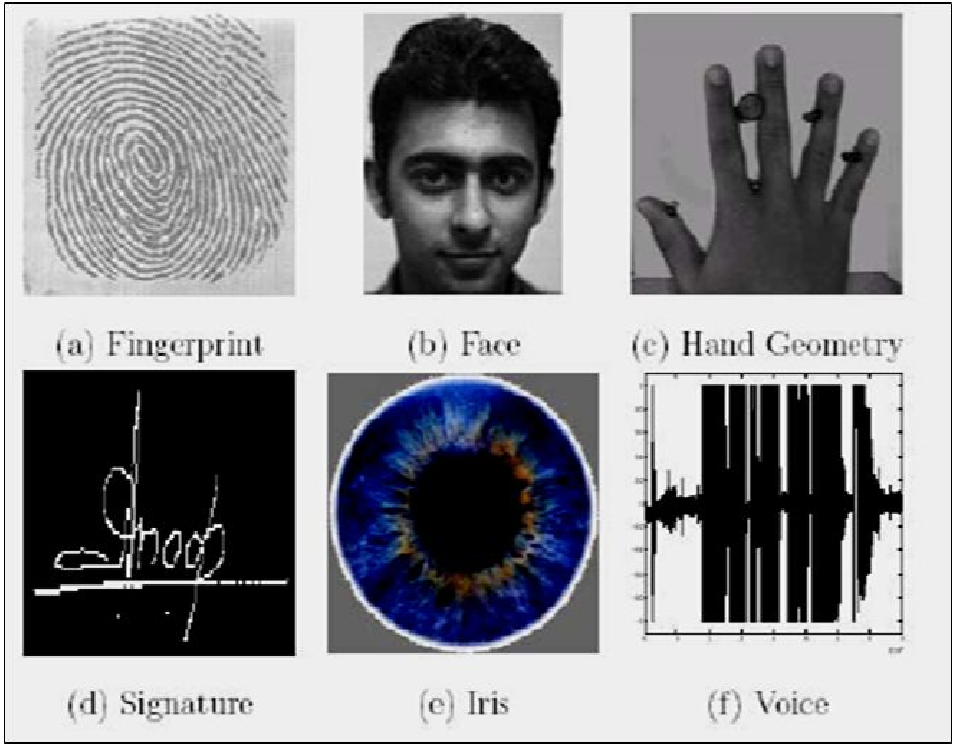
Speech Recognition



.
.
.

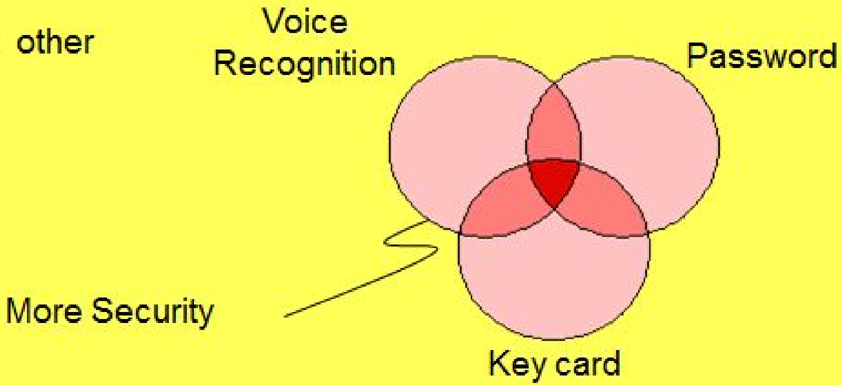
Emotion - *Happy*
Gender - *Woman*
Age - *Teengger*

Speaker ID as Biometrics



Security

- Speaker recognition may be used with other security forms to improve it. For example:
  - Password
  - Key card
  - Etc.



Applications

- Authentication.
  - Phone-based bank transactions.
  - Remote shopping.

- Data access.
  - Voice mail / internet browsing.
  - Databases.

- Access control:
  - Physical facilities.
  - Computer networks and websites.

- Law enforcement
  - Forensic applications.
  - Home parole control.

- Personalization
  - Personalization of services.

Modalities

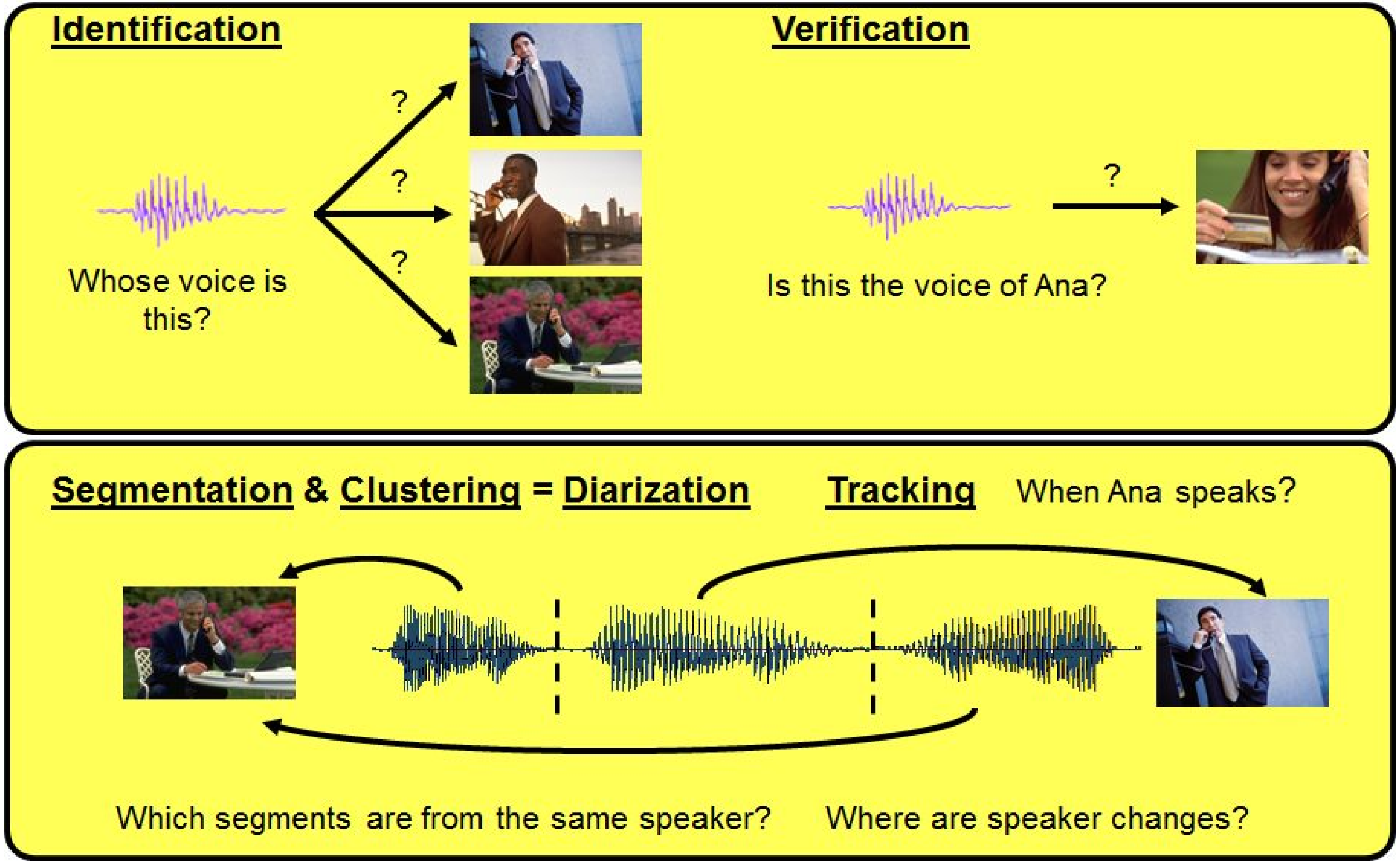
Text - dependent

- System knows the text spoken by the person.
- This knowledgment can improve performance
- It's used on systems with a high level of security requirement.

Text - independent

- System doesn't know the text spoken by the person.
- More flexible but also more difficult problem
- Speech recognition system can be used to determine the text spoken by the person.

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?

Features

Humans use different features to recognise the speaker:

- Pronunciation, diction, …
- Prosody, rhythm, speed, volume,…
- Acoustic aspects of the voice.

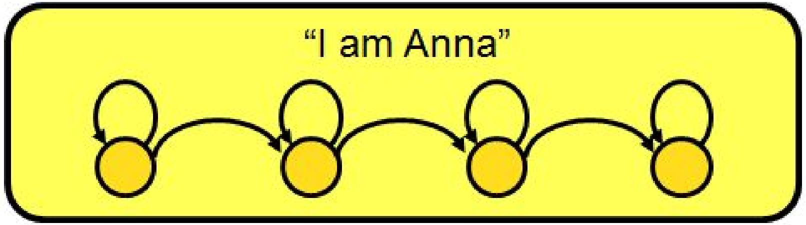
Desirable aspects for those features:

- **Practical**
  - To appear frequently and naturally during the speech
  - Easily measurable for the system
- **Robust**
  - Any change over time or by speaker’s health
  - Any change by different transmission characteristics or by background noise
- **Secure**
  - Hard to falsify

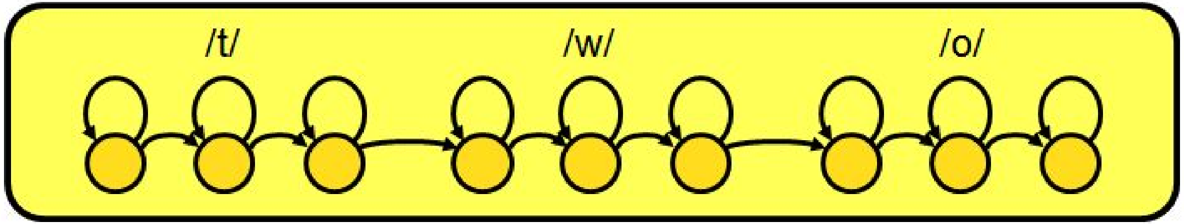
No feature has all those properties

Spectrum-derived features are the more used by now because of their effectiveness

HMMs and GMMs

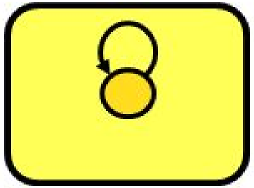


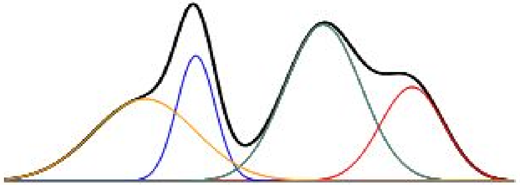
<u>Password or sentence proposed by system</u>: Phonetical model.



<u>Text independent</u>: One state model

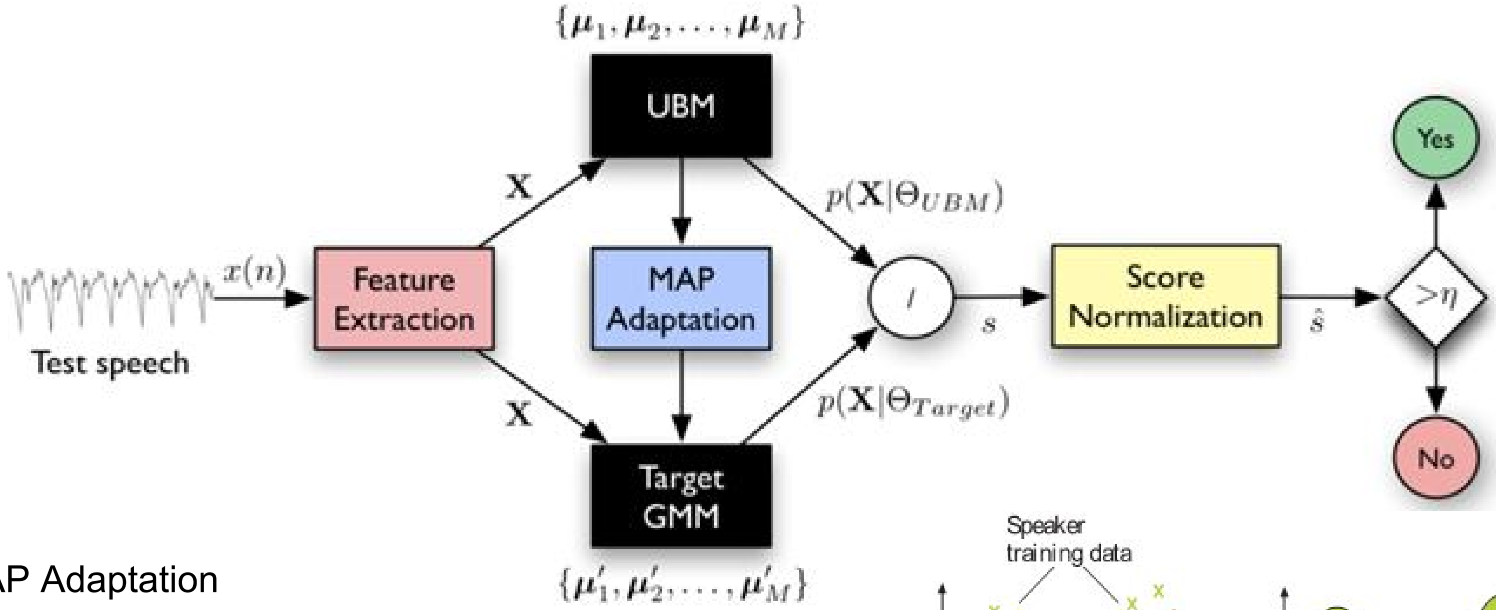
Gaussian Mixture Model (GMM)





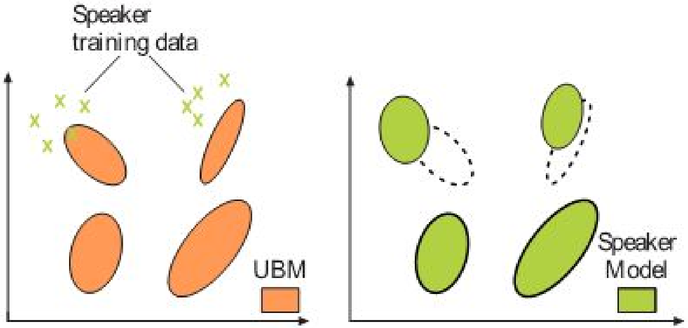
State of the art for text independent systems

GMM-UBM Universal Background Model

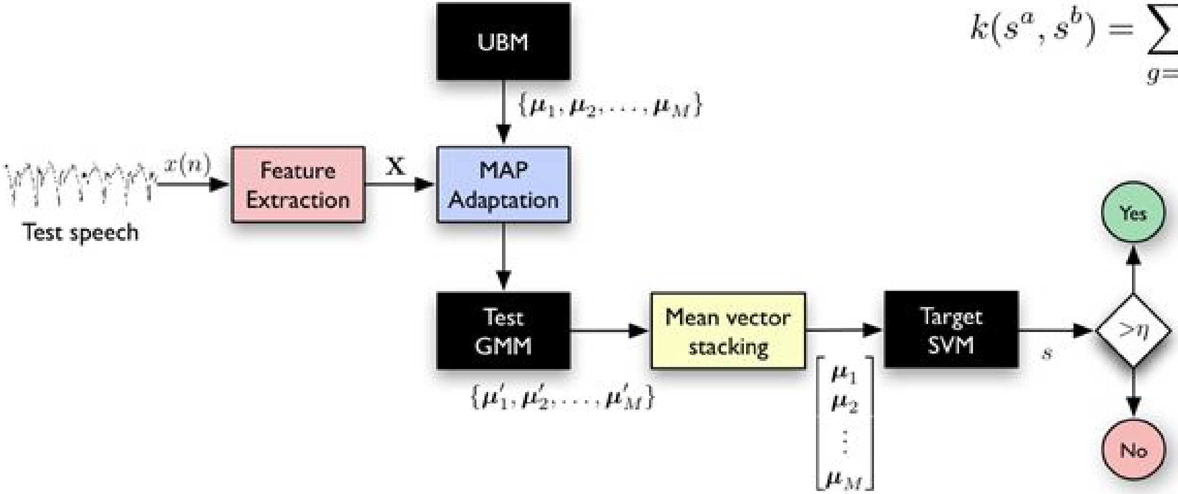


MAP Adaptation

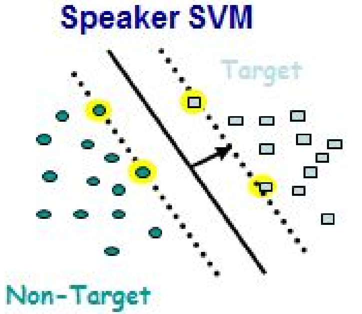
$$\mu_{client\_map} = (1-\alpha)\mu_{world} + \alpha.\mu_{client\_ML}$$



Supervectors



$$k(s^a, s^b) = \sum_{g=1}^{G} \left(\sqrt{\lambda_g}\boldsymbol{\Sigma}_g^{-\frac{1}{2}}\boldsymbol{\mu}_g^a\right)^{\mathbf{T}} \left(\sqrt{\lambda_g}\boldsymbol{\Sigma}_g^{-\frac{1}{2}}\boldsymbol{\mu}_g^b\right)$$



i-vectors

Joint Factor Analysis (JFA) model

$$s = m + Vy + Ux + Dz$$

"Ideal" speaker supervector

Speaker-independent component

Speaker-dependent component

Channel-dependent component

Speaker-dependent residual component

$$s = m + Tw$$

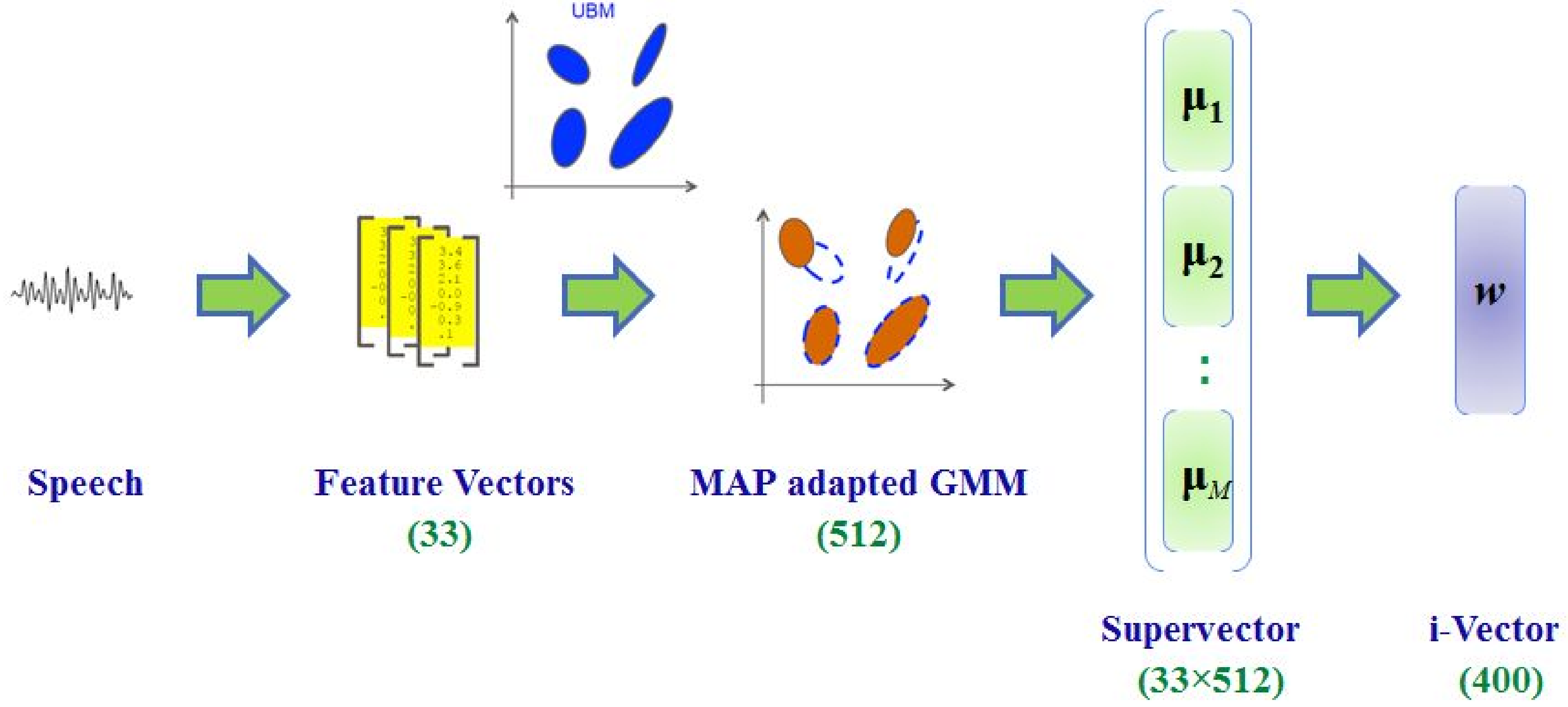
Speaker Supervector

Speaker-independent Component

Total-variability Matrix

i-vector

i-Vector dimension



i-Vector Training

T is trained iteratively according to the GMM posteriors

0th order statistic $\longrightarrow N_c(s) = \sum_{t \in s} \gamma_t(c)$

The posterior of Gaussian component c for observation t of speaker s

Sum over feature frames of all relevant conversation sides of speaker s

1st order statistic $\longrightarrow F_c(s) = \sum_{t \in s} \gamma_t(c) Y_t$

2nd order statistic $\longrightarrow S_c(s) = diag\left(\sum_{t \in s} \gamma_t(c) Y_t Y_t^* \right)$

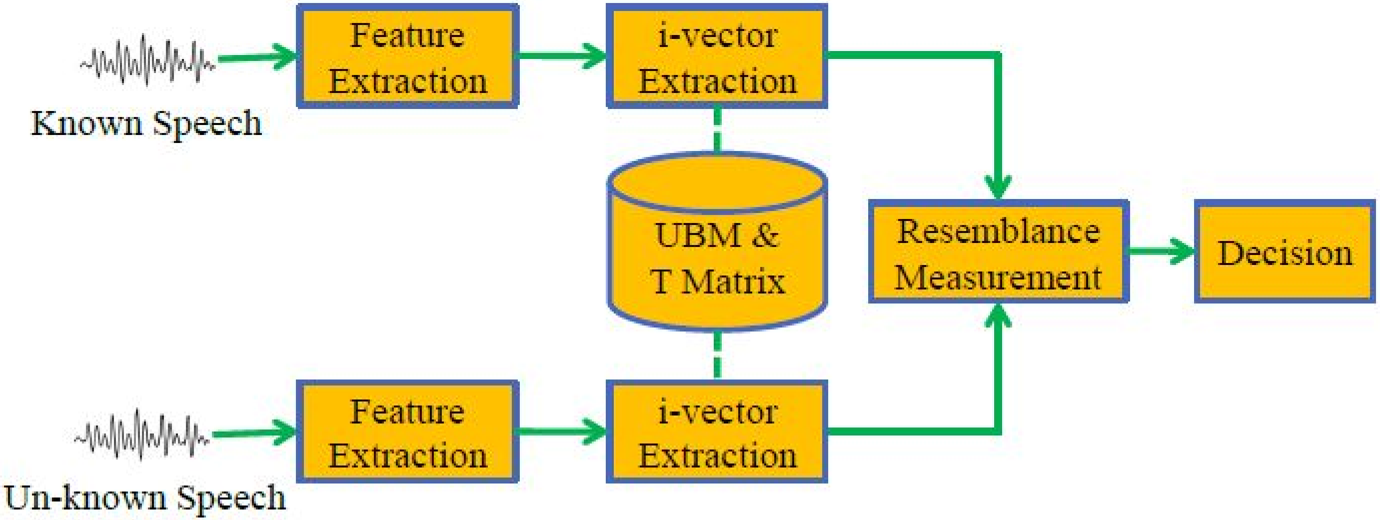
Denotes Hermitian transpose of vector or matrix

Keep only the diagonal entries; zero out other entries

Given T and the UBM, i-vectors are extracted for each speaker utterance

$$w = \left(I + T^t \Sigma^{-1} N(u) T\right)^{-1} . T^t \Sigma^{-1} \tilde{F}(u).$$

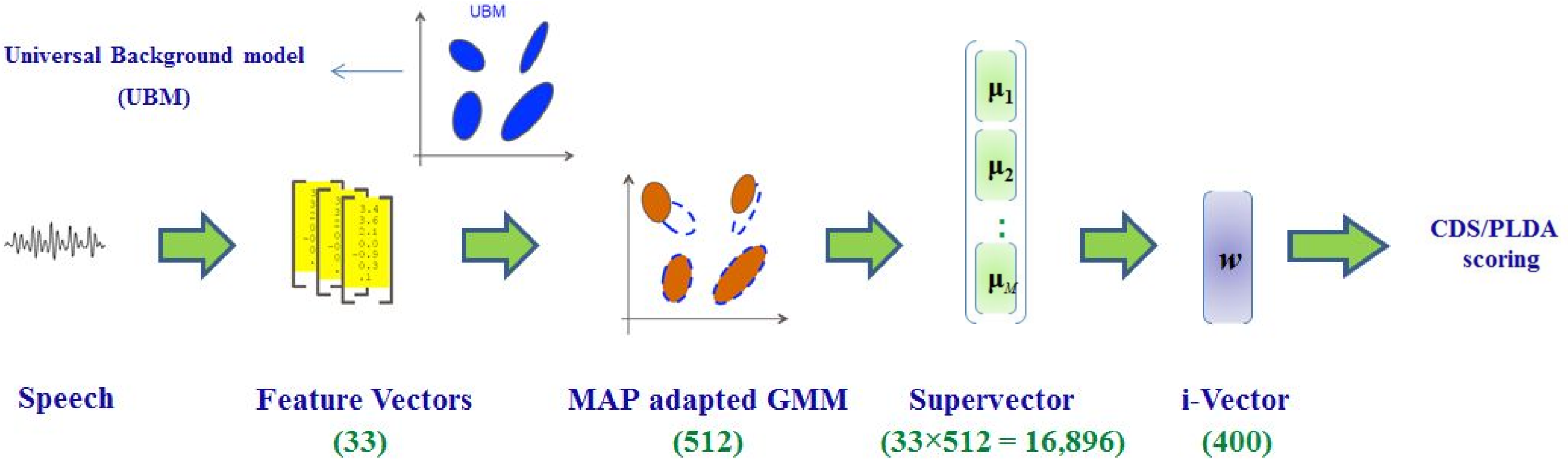
i-vector Scoring



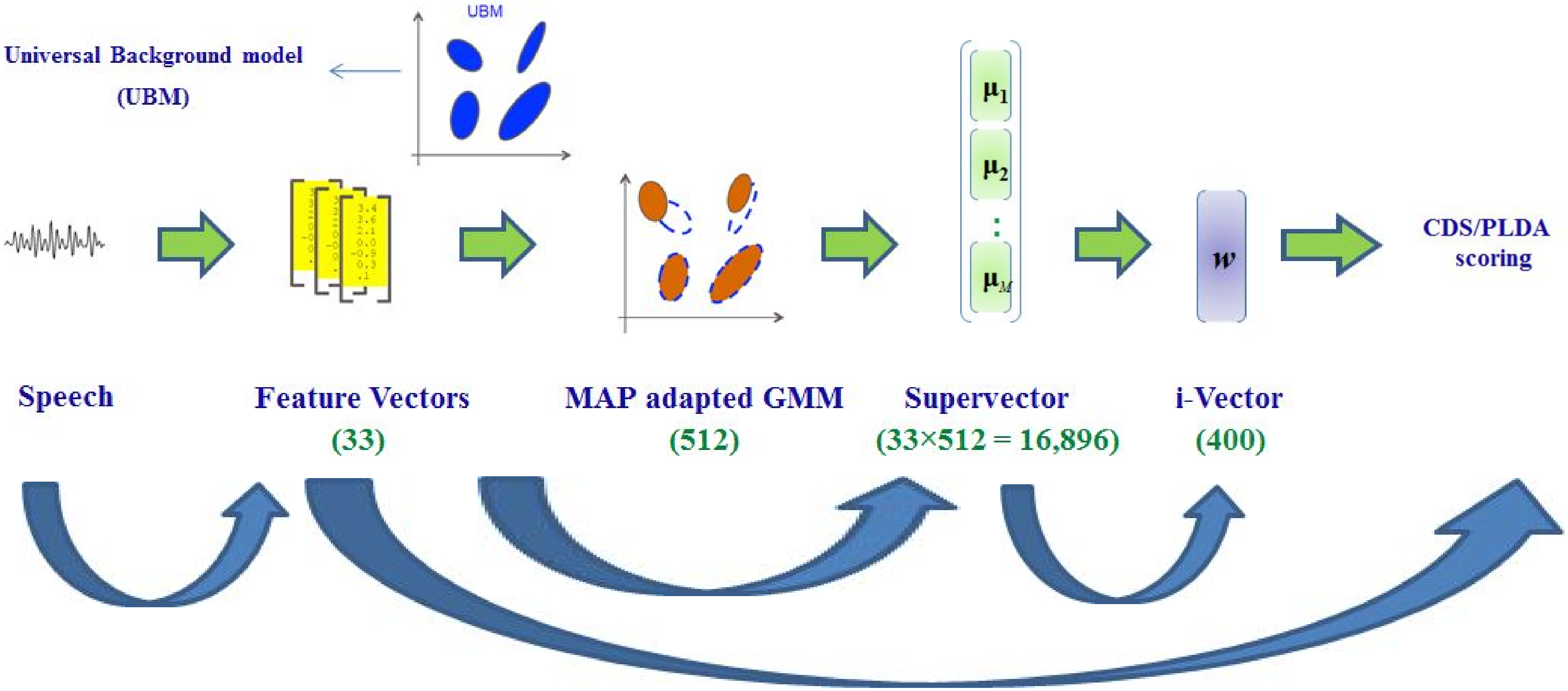
Resemblance Measurement

- Cosine Distance Scoring $$score(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1^T . \mathbf{w}_2}{\|\mathbf{w}_1\| . \|\mathbf{w}_2\|} = cos(\theta_{\mathbf{w}_1, \mathbf{w}_2})$$
- Probabilistic Linear Discriminant Analysis (PLDA)

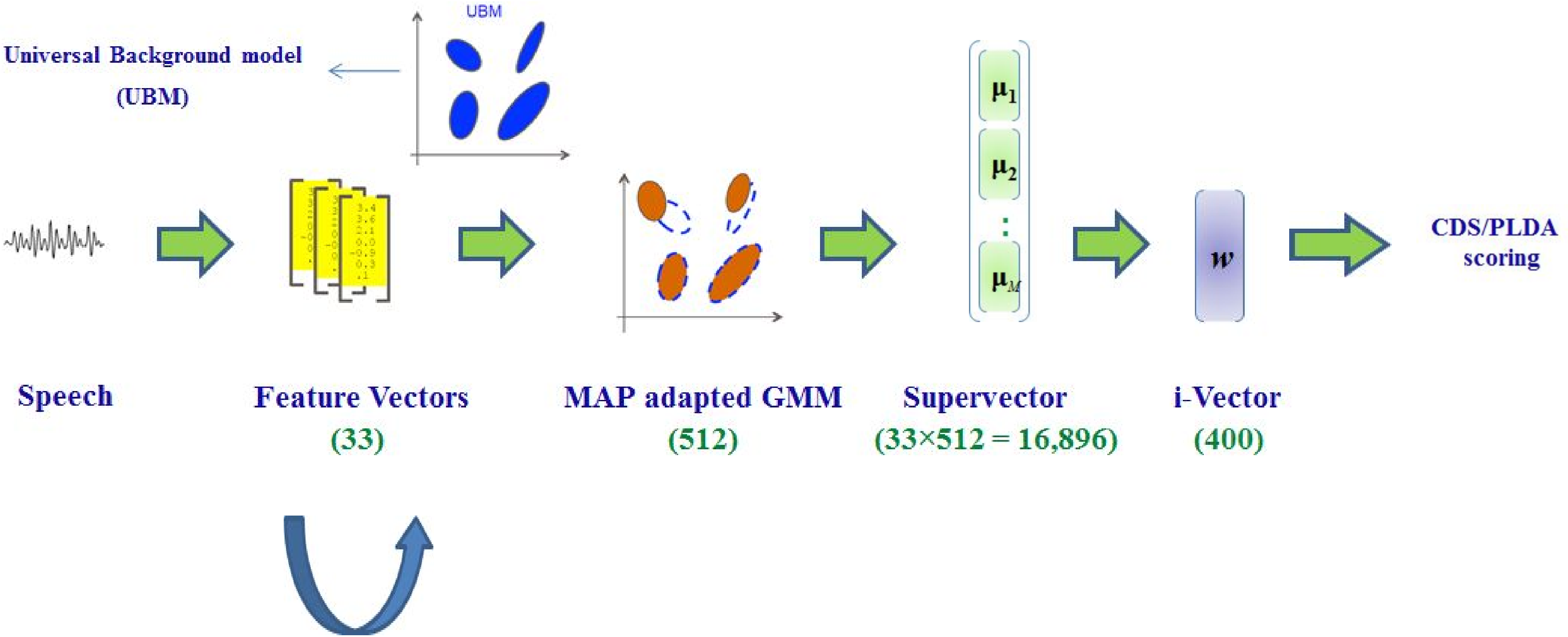
SoA Speaker Recognition



DL in Speaker Recognition

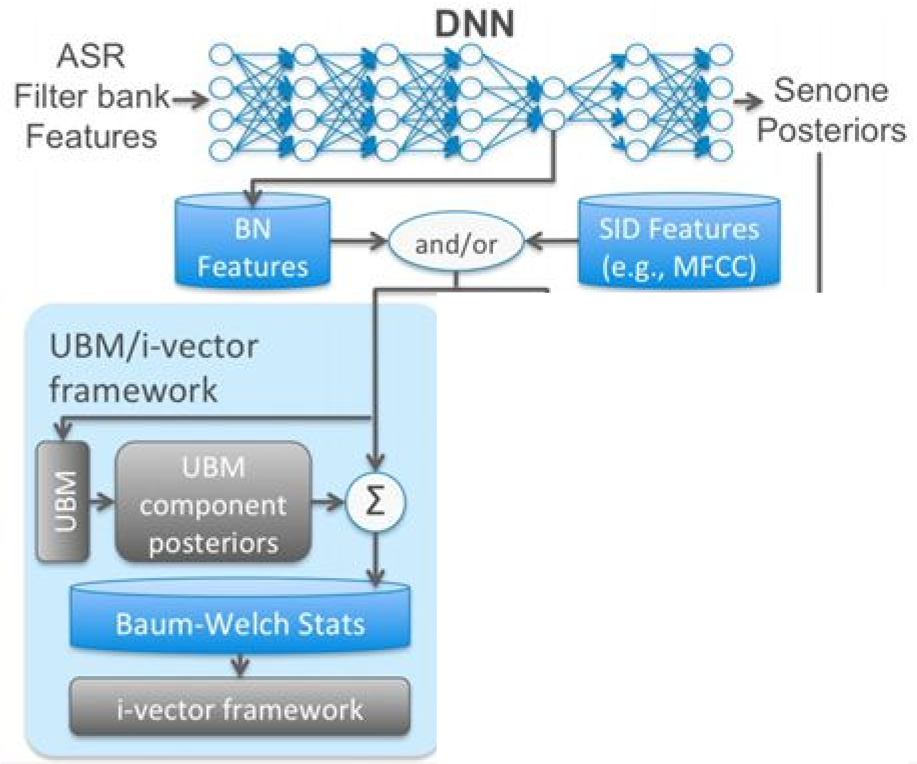


DL Features



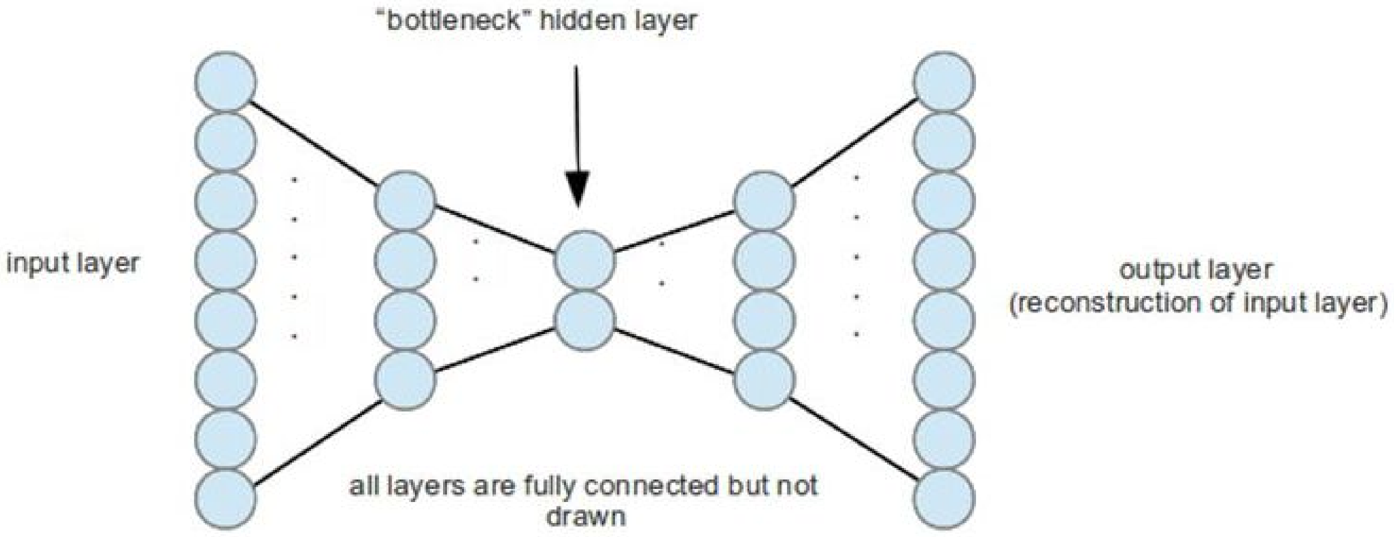
Universal Background model
(UBM)
UBM
μ_1
μ_2
μ_M
w
CDS/PLDA scoring
Speech
Feature Vectors
(33)
MAP adapted GMM
(512)
Supervector
(33×512 = 16,896)
i-Vector
(400)

BN Features



After M. Mclaren e al.,“Advances in deep neural network approaches to speaker recognition” ICASSP 2015.

Autoencoder

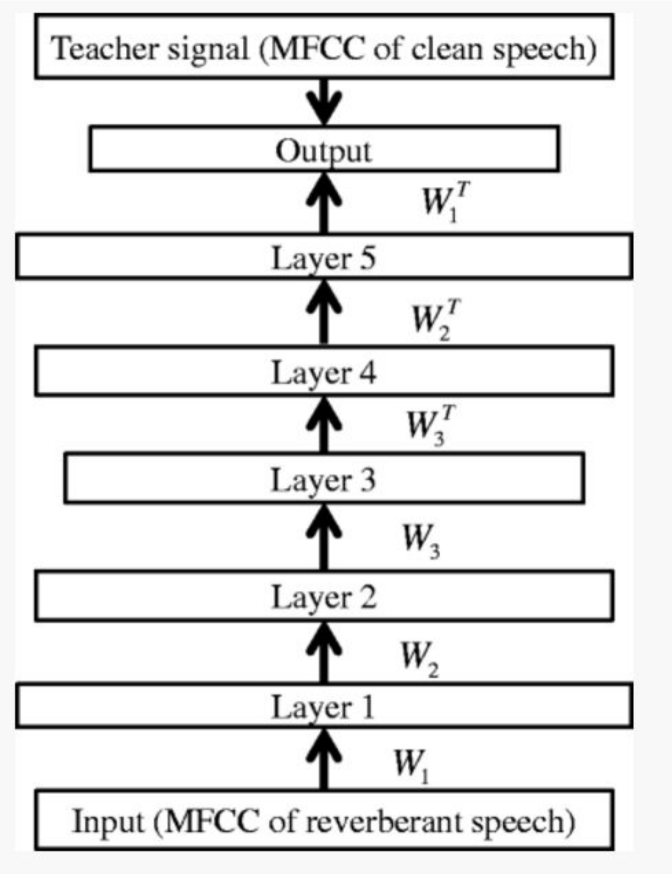


Denoising Autoencoder

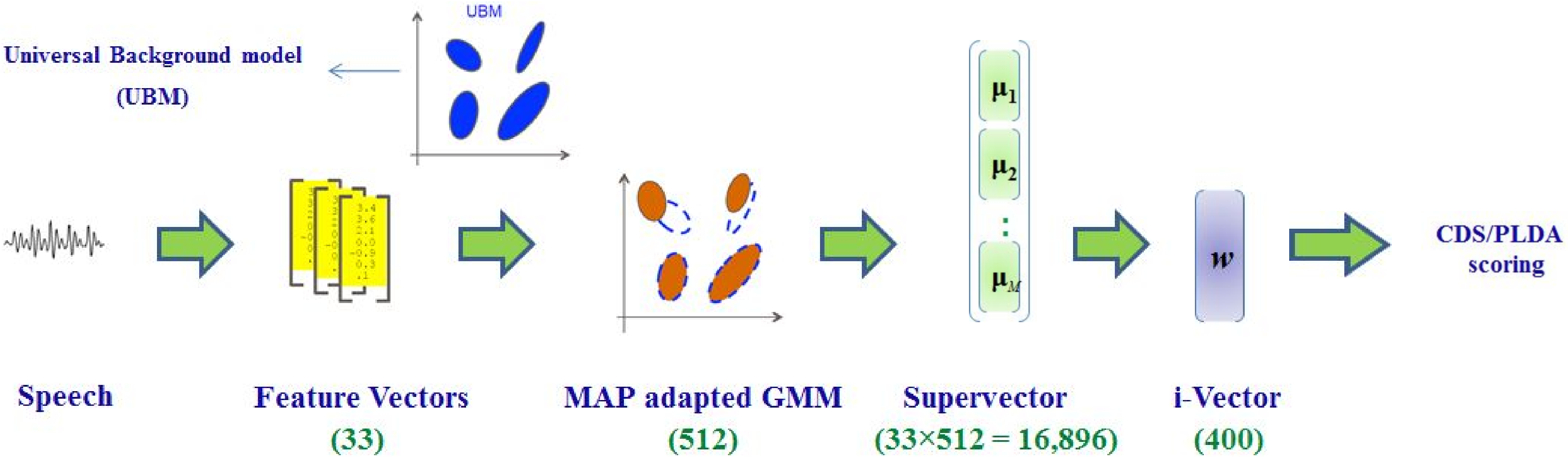
Denoising autoencoder for cepstral domain dereverberation.

- Transfrom noisy features of reverberant speech to clean speech features.
- Pre-Trainning with Deep Belief Networks (DBN)

Zhang et al., Deep neural network-based bottleneck feature and denoising autoencoder-based fro distant-talking speaker identification, EURASSIP Journal on Audio, Speech, and Music Processing (2015) 2015:12



DL Modeling i-Vectors



14
omid

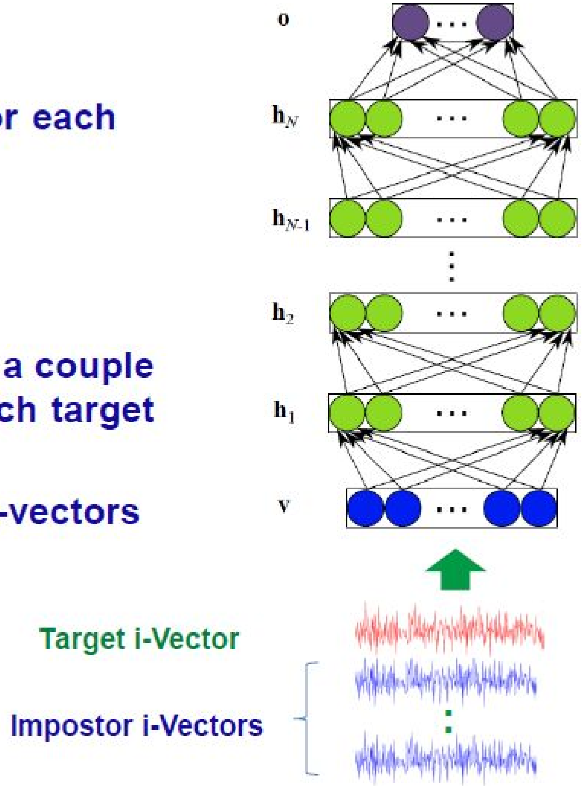
DL Modeling i-Vectors

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)



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

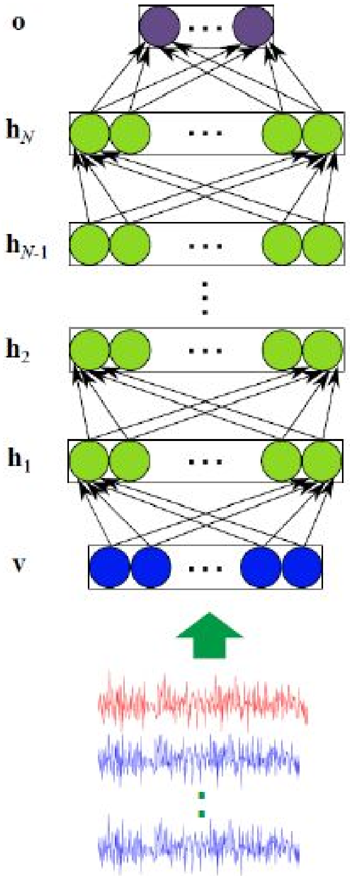
DL Modeling i-Vectors

Problems :

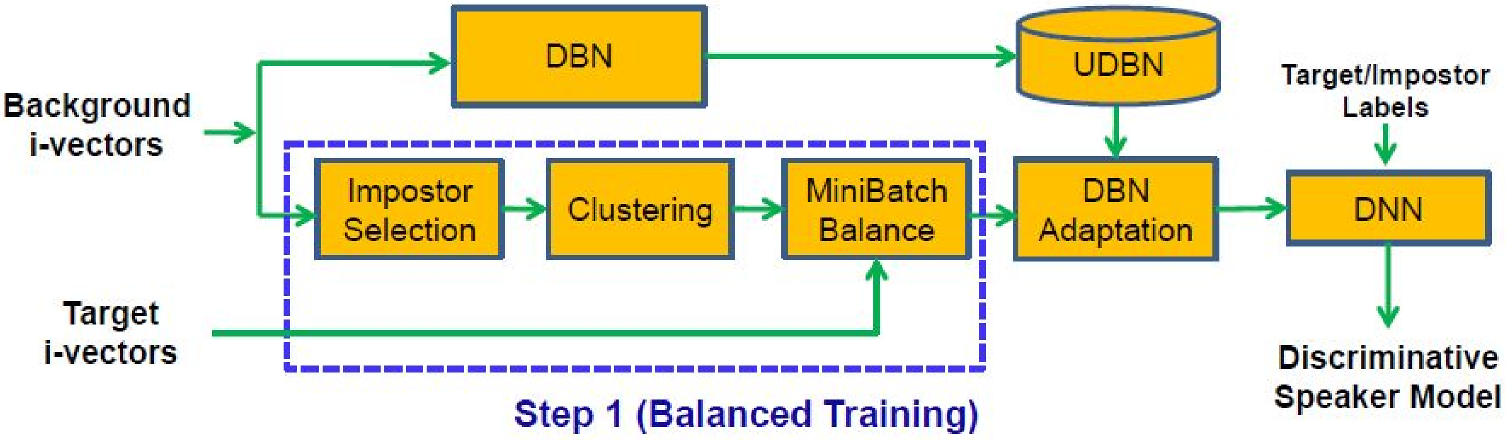
- Unbalanced data → Bias towards the majority class
- Few data → Overfitting

Our Proposal:

- **Balanced training**
  - ✓ Impostor selection and clustering
  - ✓ Distributing equally impostor and target samples among minibatches
- **DBN Adaptation**
  - ✓ Take advantage of unsupervised learning of DBN using the whole background data called Universal DBN (UDBN)
  - ✓ Adapt UDBN to few data of each speaker



Decoder



Step 1 : Balanced Training

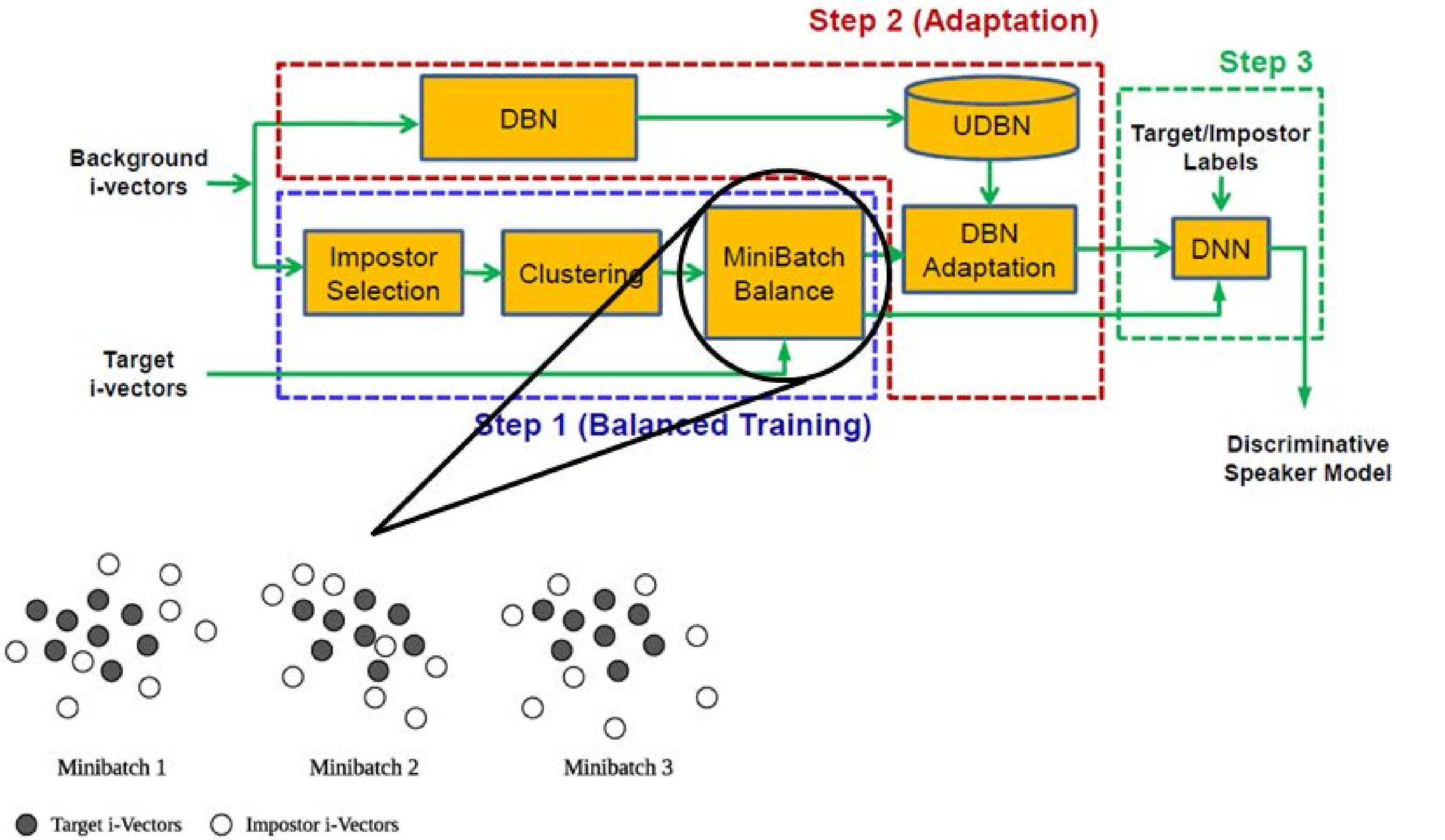
Problem:

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

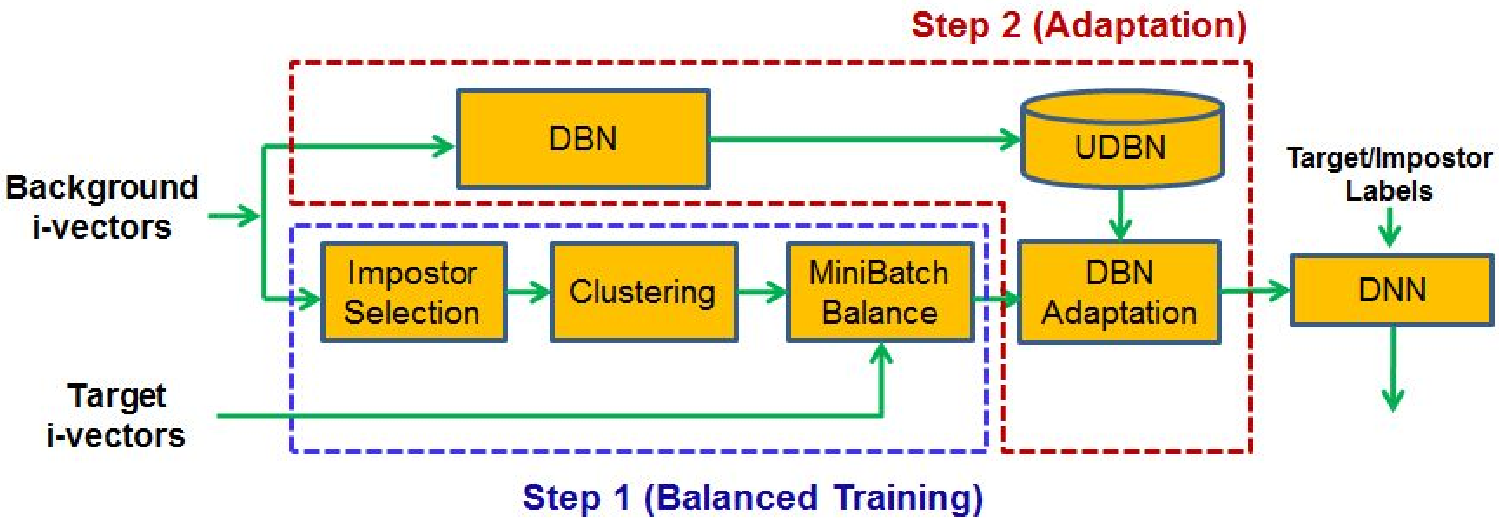
Solutions:

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

DL Modeling i-Vectors



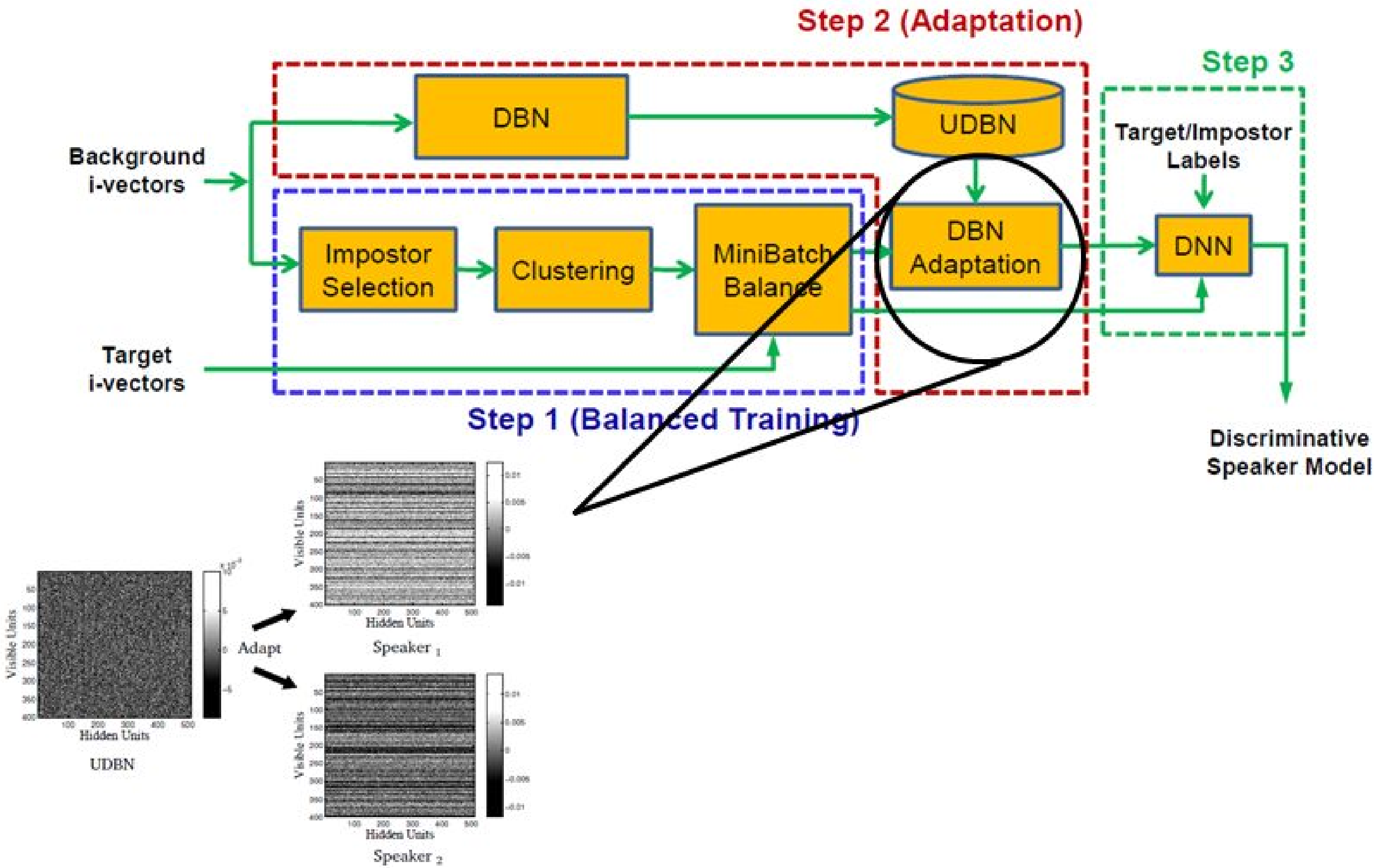
DL Modeling i-Vectors



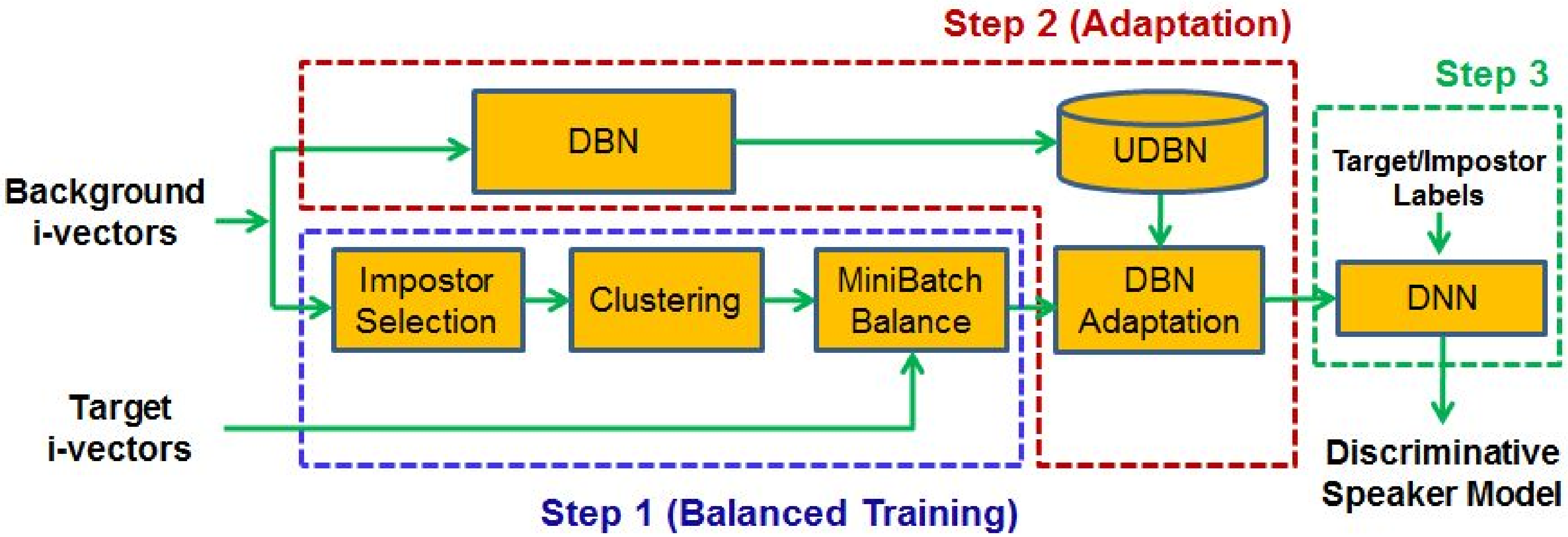
Step 2 : Adaptation

- 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

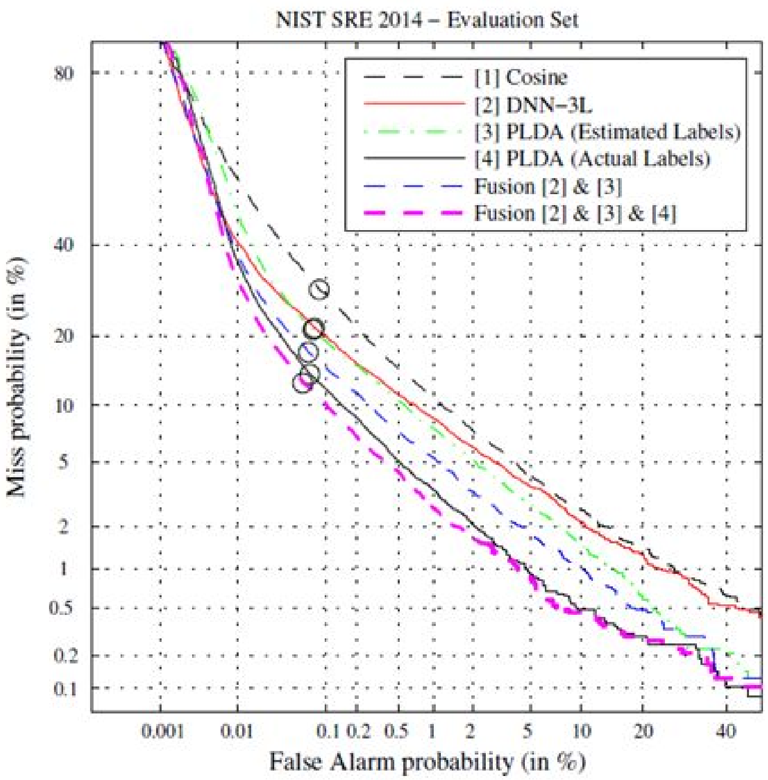
- Supervised learning given impostor and target labels, adapted DBN, and balanced data

DL Modeling i-Vectors

NIST SRE 2014 i-Vector Challenge

- 59,729 telephone calls from 6,087 speakers
- Speech signals with variable durations following a long-normal distribution with the mean of 39.6 sec
- 1,306 target speakers
- 12,582,004 trials
- 40% of data as the progress set and 60% as the evaluation
- 600 dimensional i-vectors

DL Modeling i-Vectors



	Labeled Background Data	Prog Set		Eval Set			
		EER	minDCF	EER	minDCF		
[1] Cosine	No	4.78	386	4.46	378	23%	37%
[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	6%	11%
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**		

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)