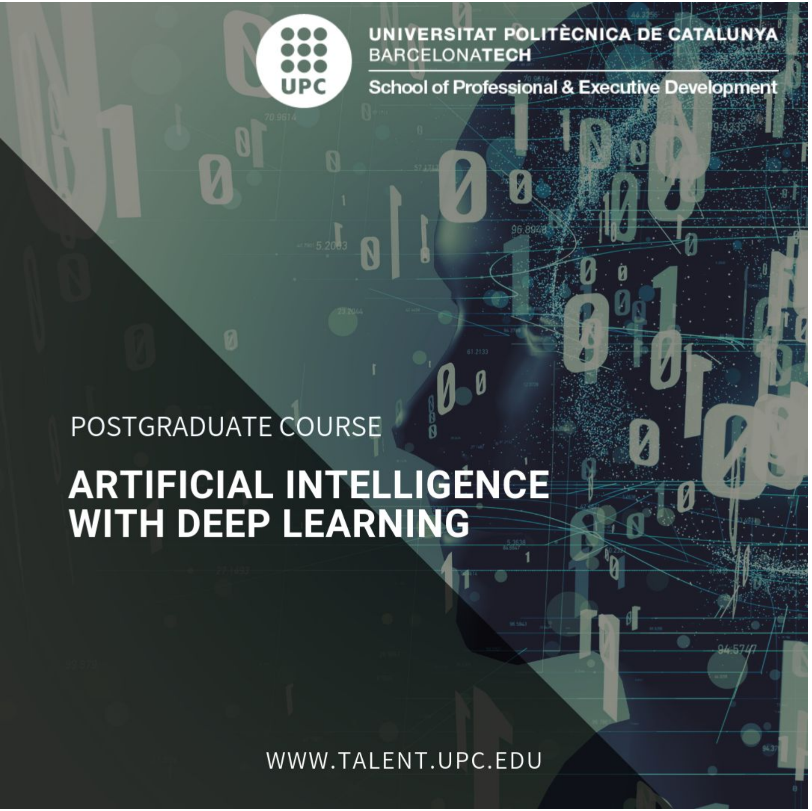


#DLUPC

Speech and Audio Processing

Speaker Recognition

Javier Hernando
javier.hernando@upc.edu

Full Professor
Universitat Politecnica de Catalunya
Technical University of Catalonia

Acknowledgments

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

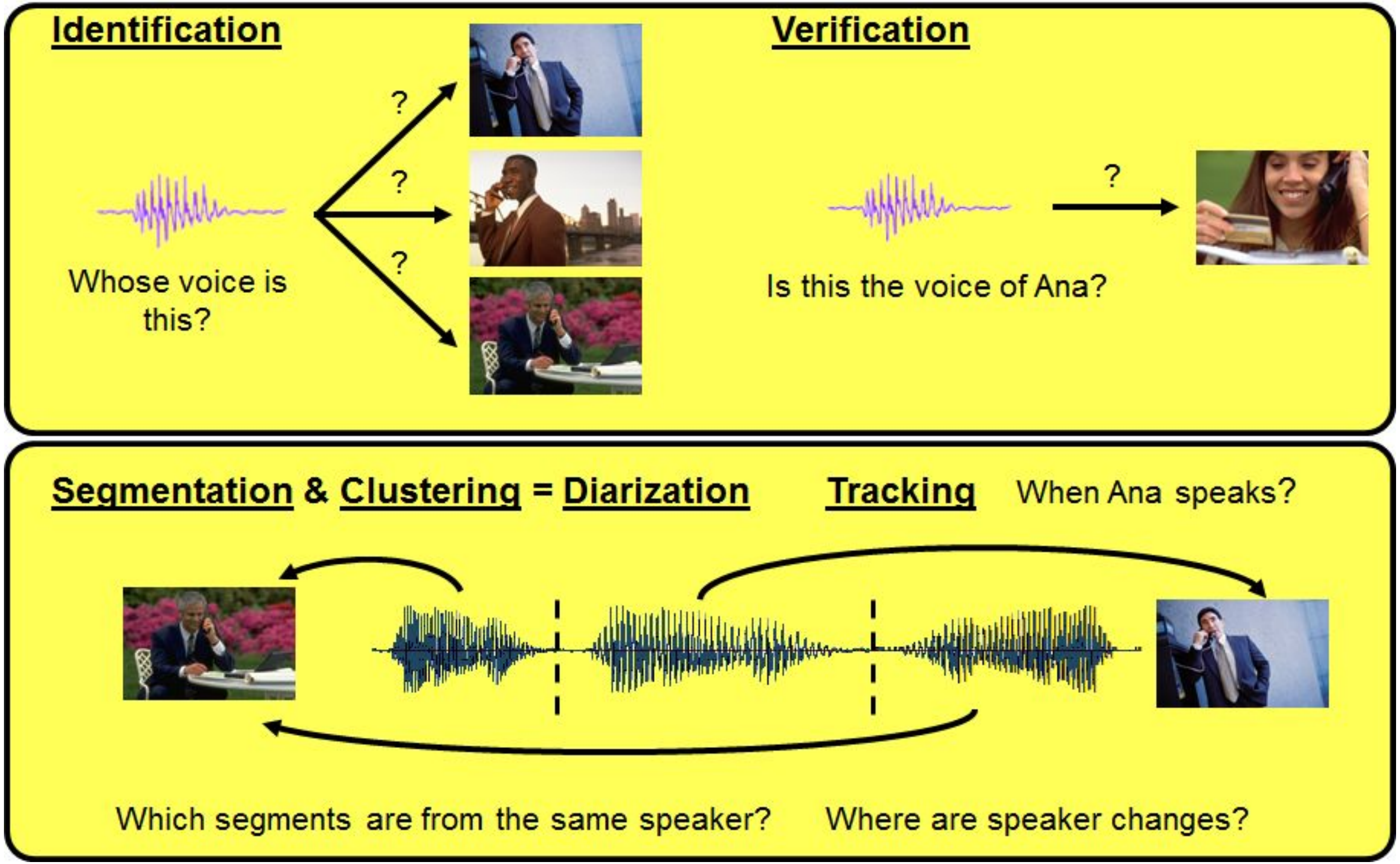
UPC TALP

TALP Research Center

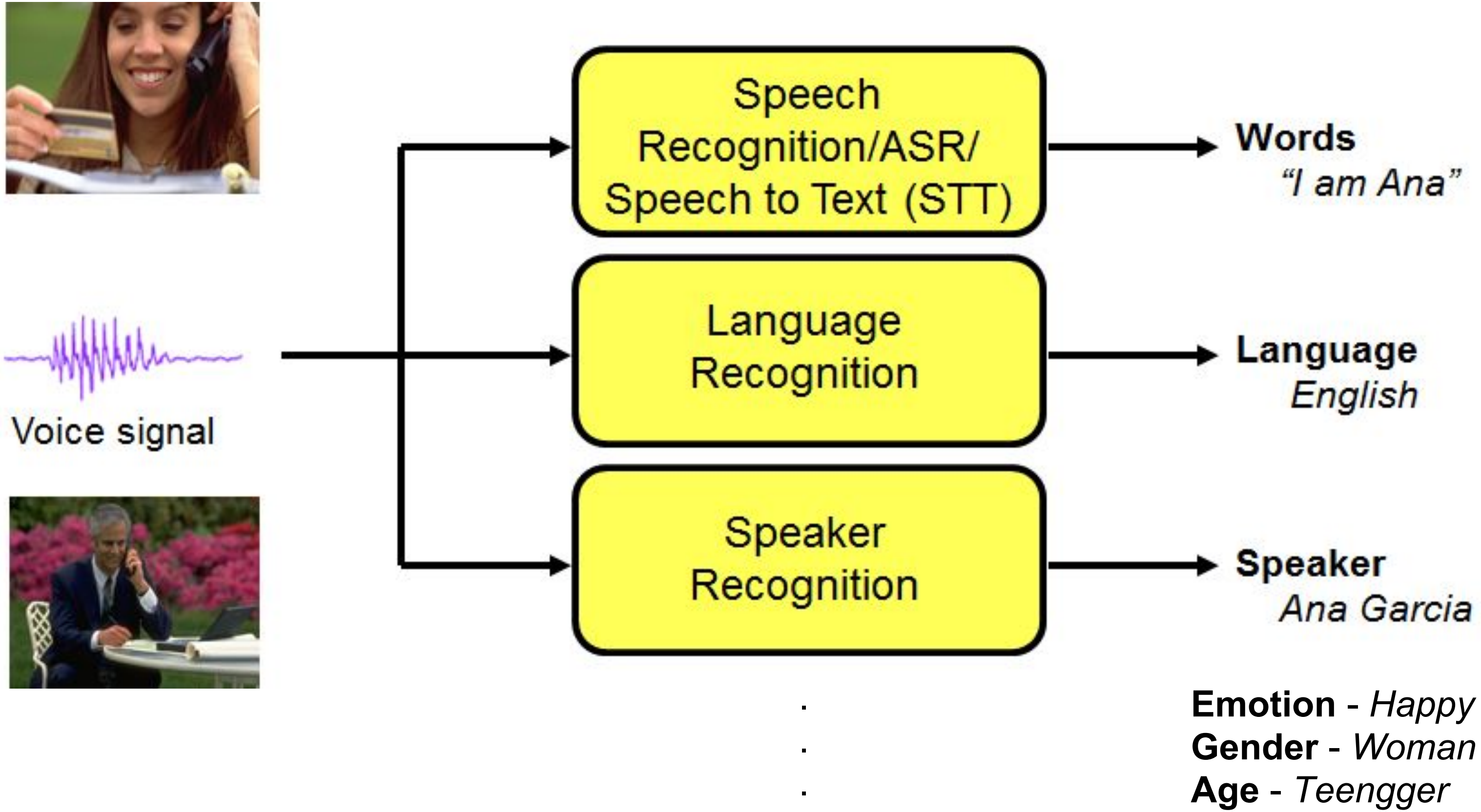
Outline

- State-of-the-art Speaker Recognition
- DL in Speaker Recognition
    - End-to-End
    - Front-End
        - Features
        - i-vector Extraction
        - Features to Embeddings
        - Vectors to Embeddings
    - Back-End

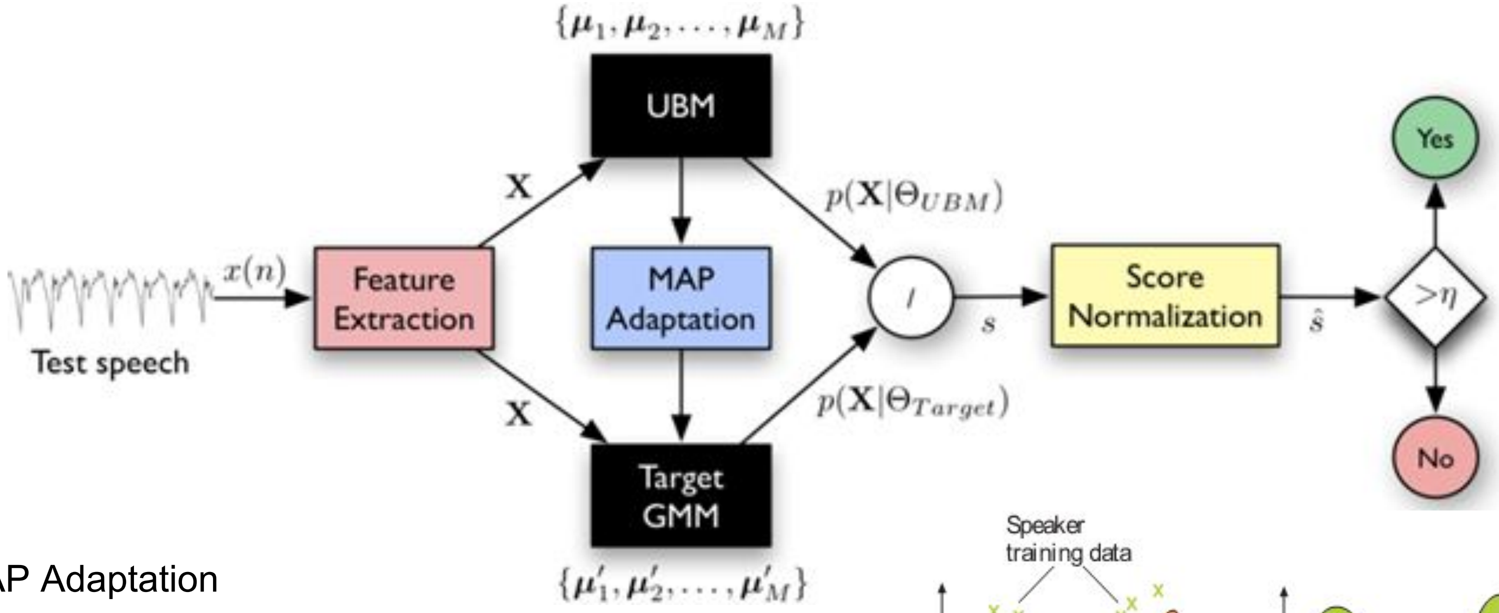
Speaker Recognition Tasks



Speech Recognition Tasks

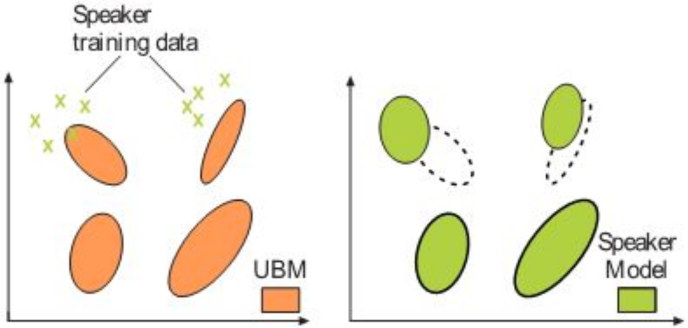


GMM-UBM Universal Background Model



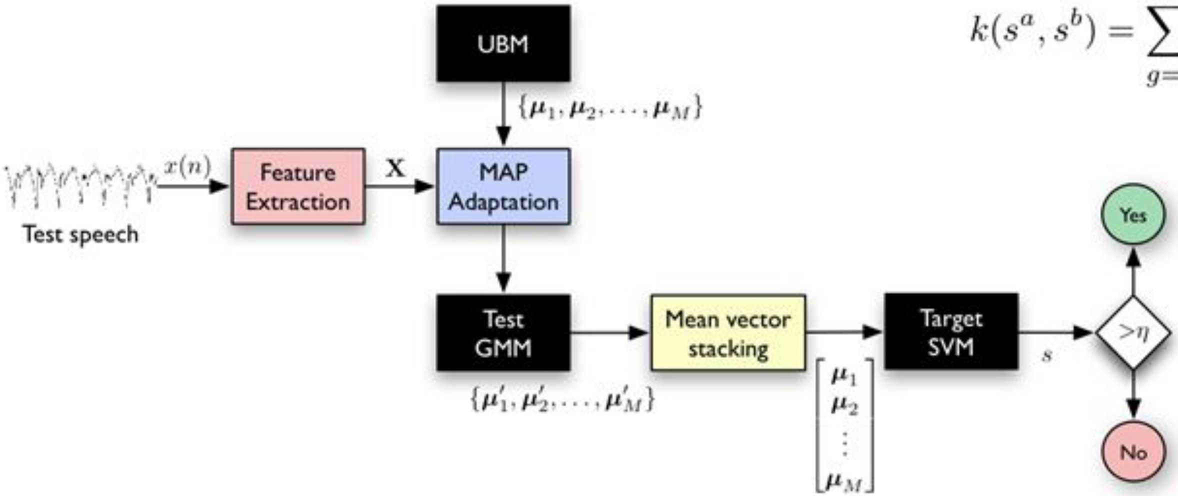
MAP Adaptation

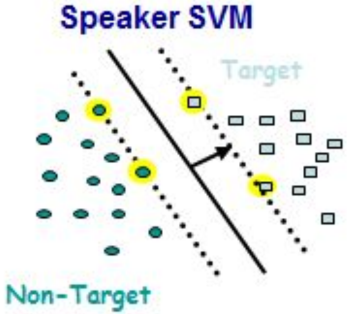
$$\mu_{client\_map} = (1-\alpha)\mu_{world} + \alpha.\mu_{client\_ML}$$



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

$$s = m + Tw$$

Speaker Supervector

Speaker-independent Component

Total-variability Matrix

i-vector

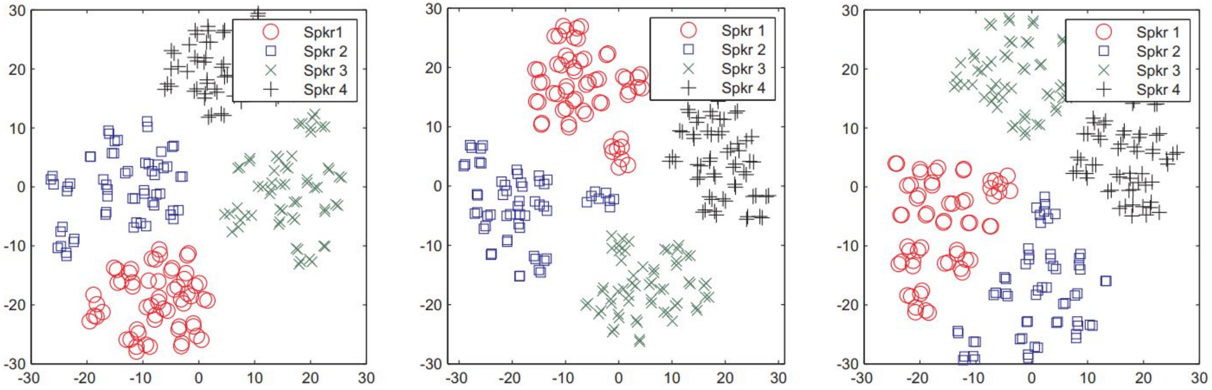
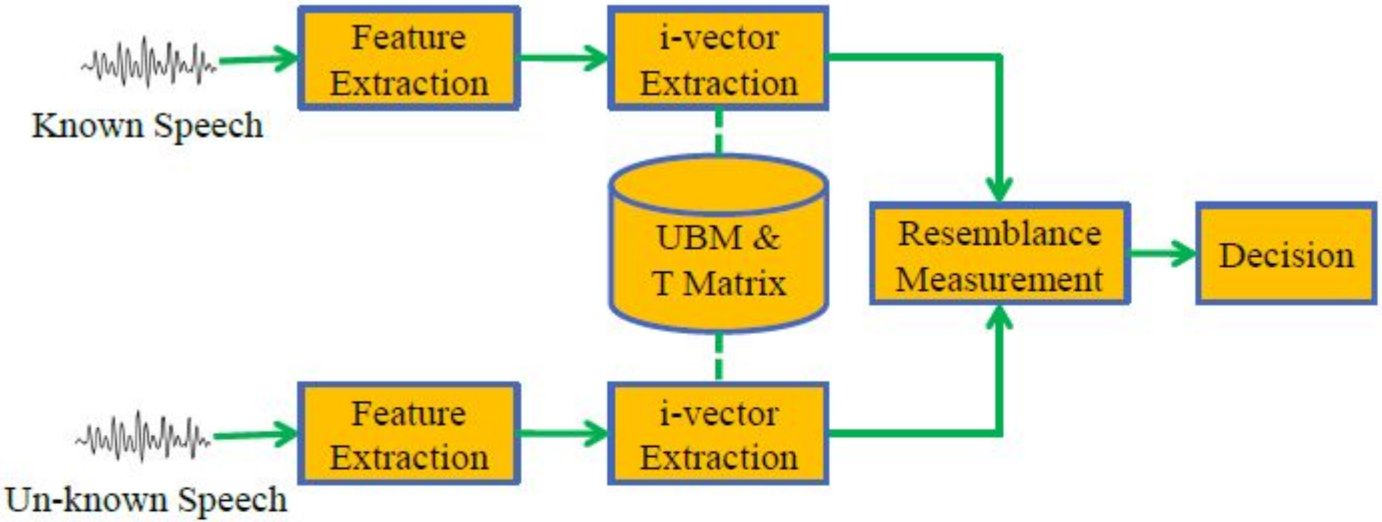


Fig. 6. t-SNE visualization of i-vectors obtained from speaker verification system using (a) IFCC, (b) FDLP and (c) MFCC features.

K. Vijayan et al./Speech Communication 81 (2016) 54–71

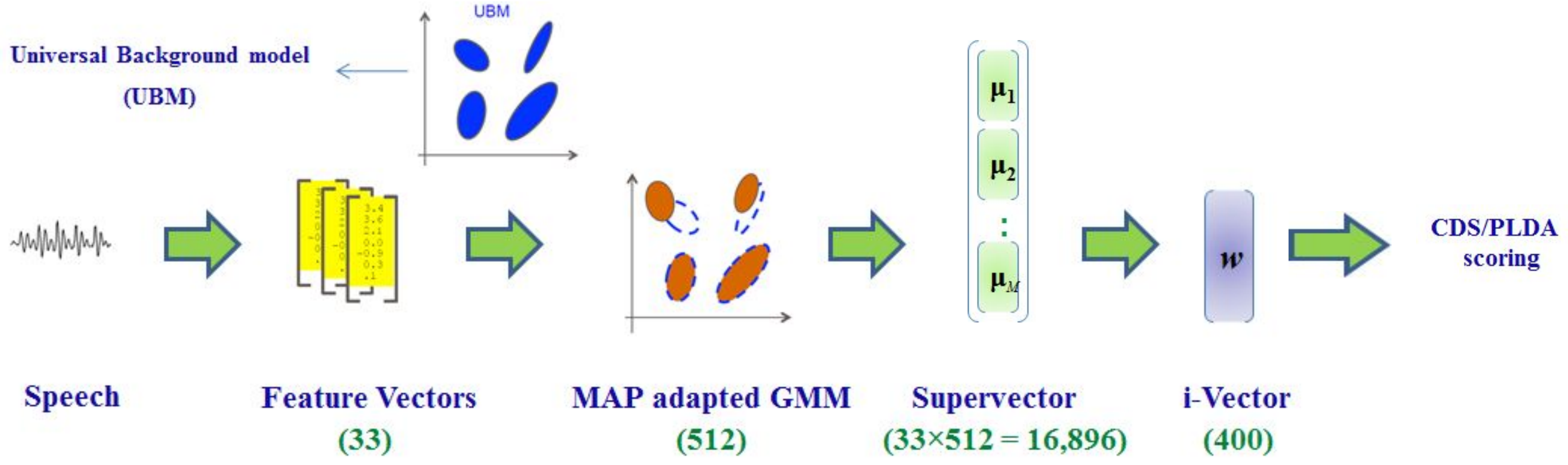
i-vector Scoring



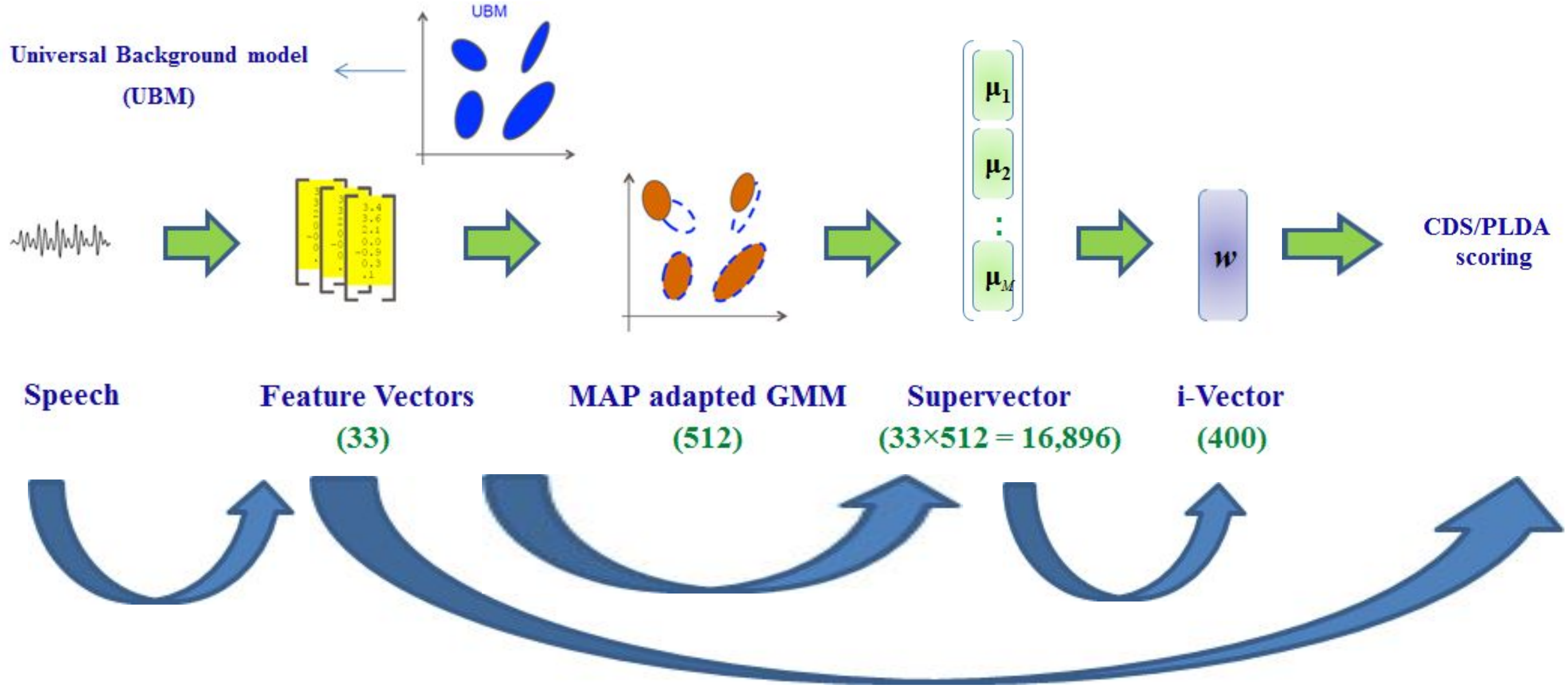
Resemblance Measurement

- 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)

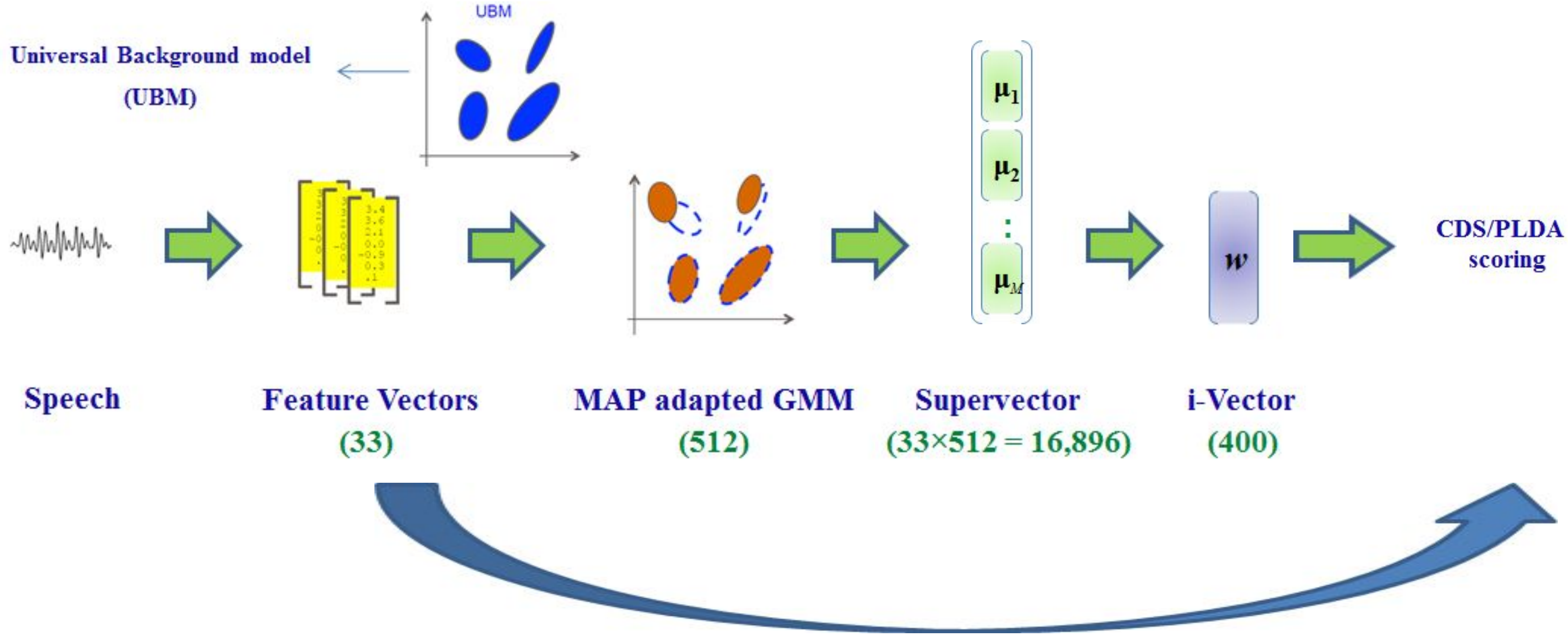
SoA Speaker Recognition



DL in Speaker Recognition

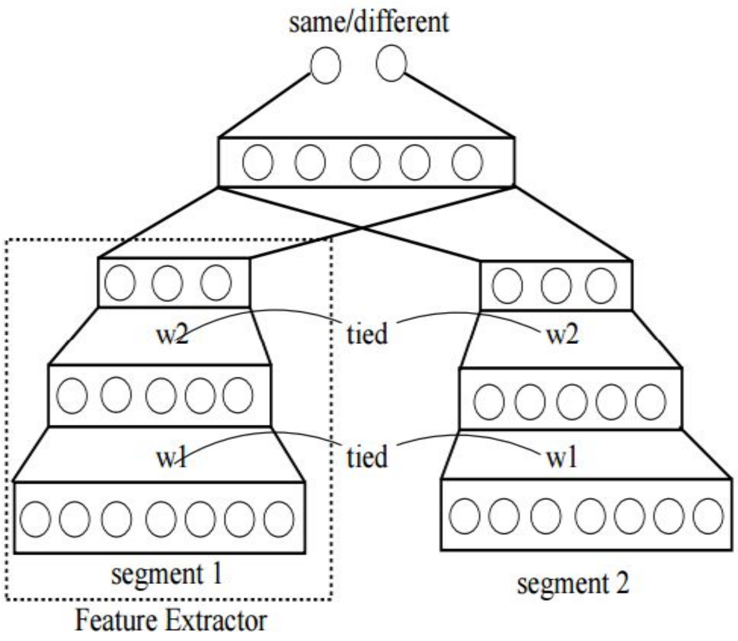


End-to-End

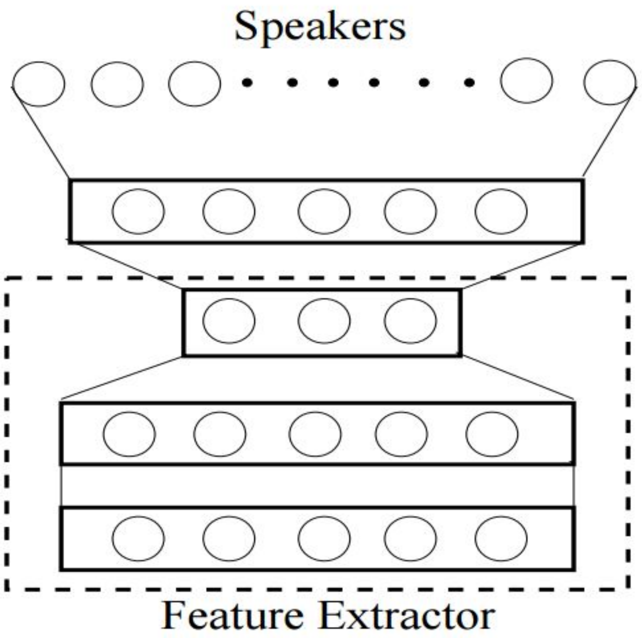


UBM
Universal Background model
(UBM)
μ_1
μ_2
μ_M
w
CDS/PLDA
scoring
Speech
Feature Vectors
(33)
MAP adapted GMM
(512)
Supervector
(33×512 = 16,896)
i-Vector
(400)

End-to-End



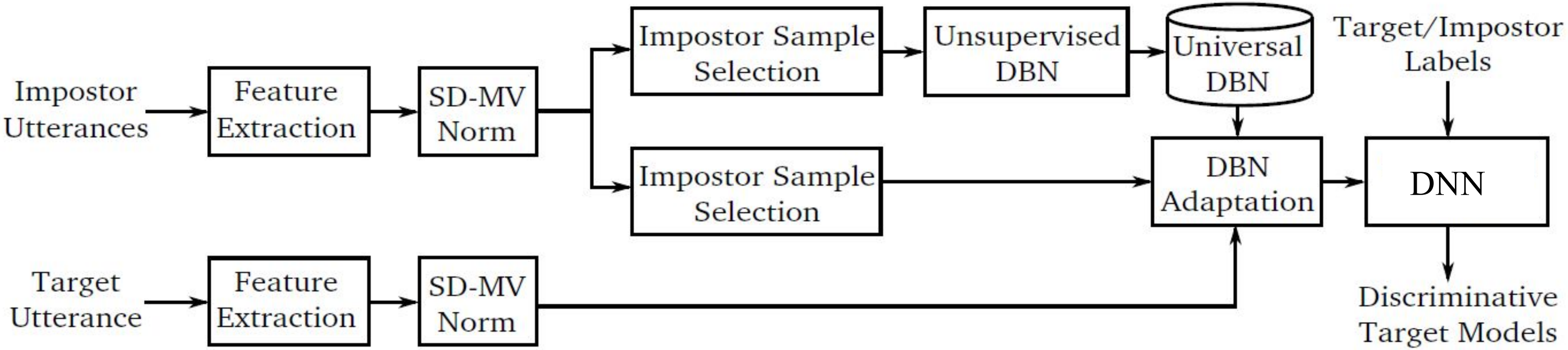
Speaker Verification



Speaker Identification

Yella, S.H., Stolcke, A., "A comparison of neural network feature transforms for speaker diarization." Proc. *INTERSPEECH* 2015.

ADBN Feature Classification



P. Safari, O. Ghahabi, J. Hernando, “Feature classification by means of Deep Belief Networks for speaker recognition”, Proc. EUSIPCO 2015

ADBN Feature Classification

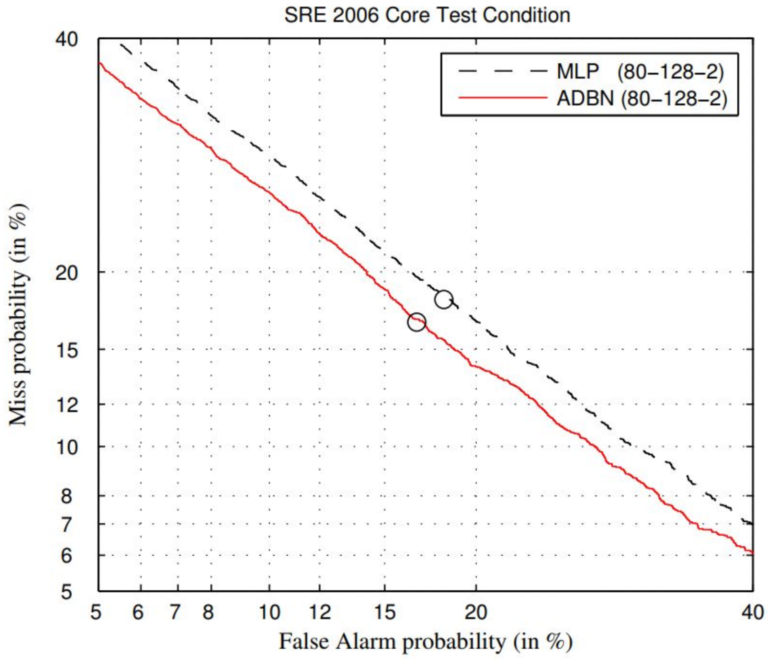
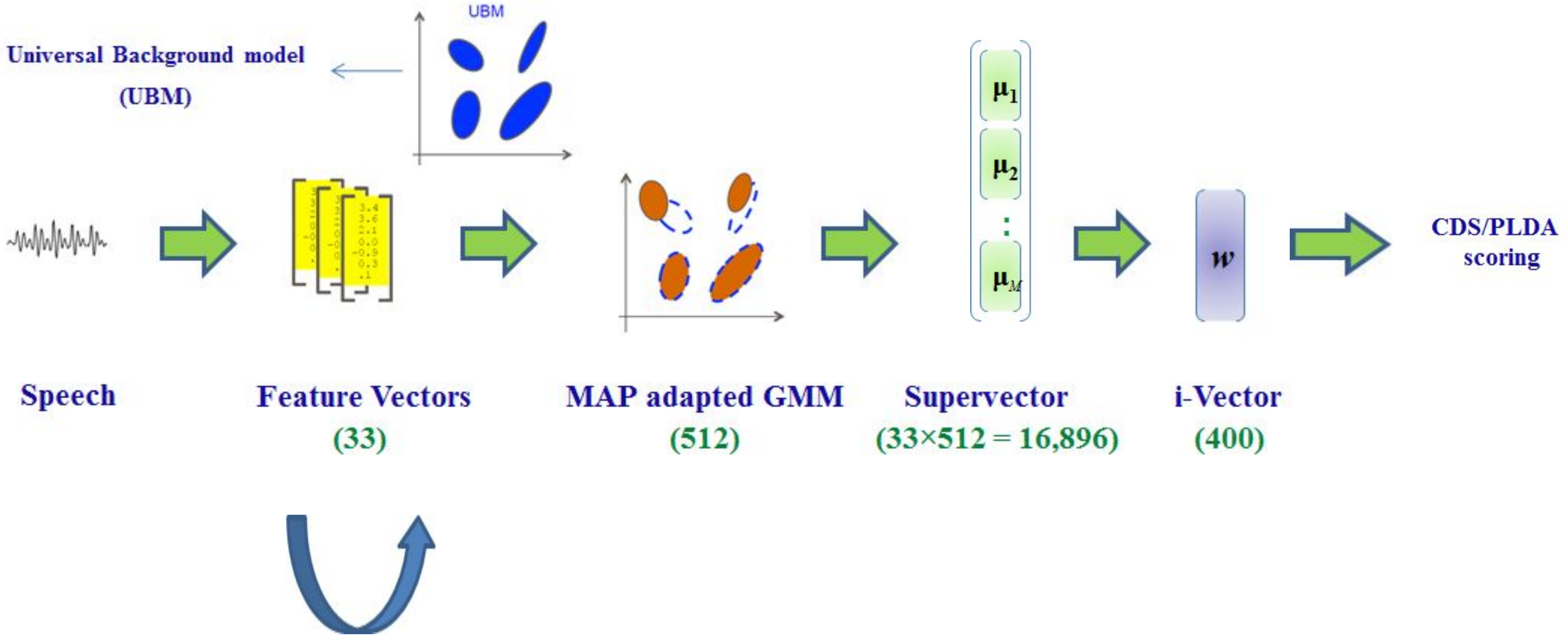
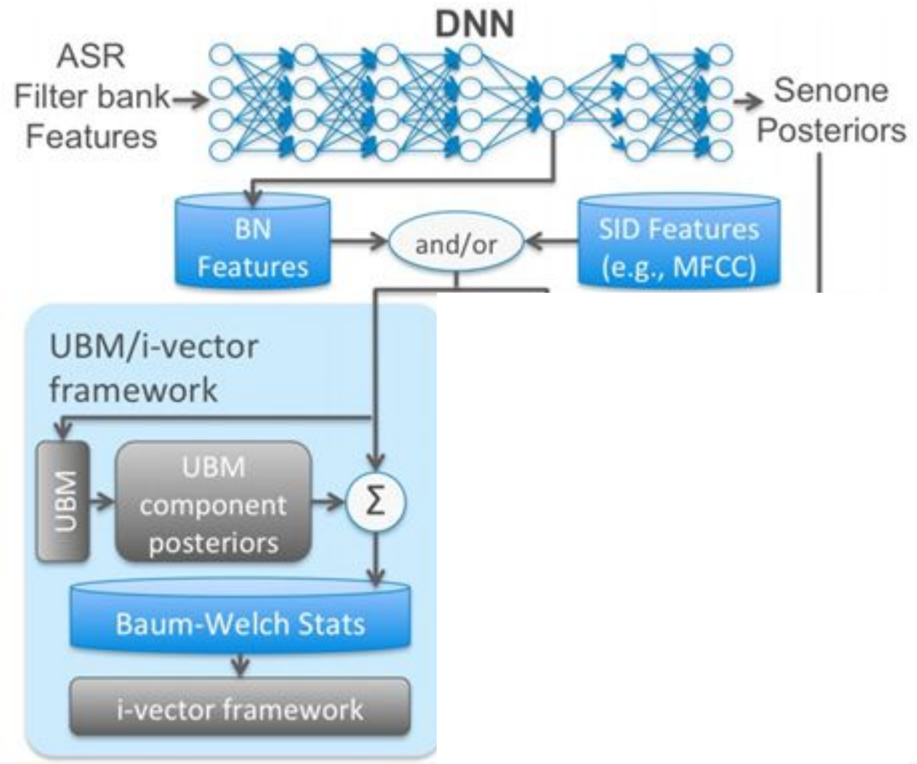


Fig. 5. Comparison of DET curves for MLP and the proposed ADBN.

Front-End: Features



ASR BN Features



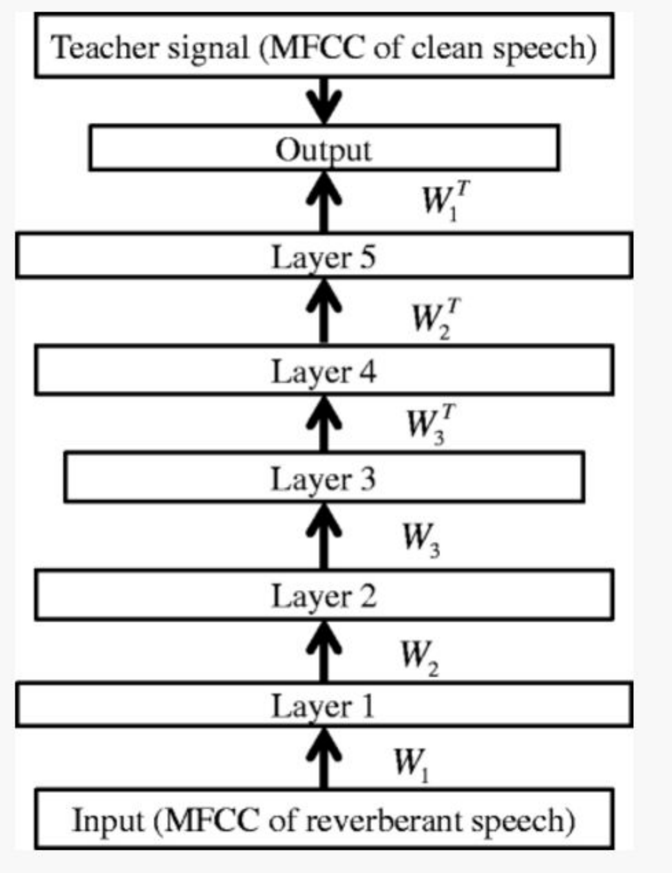
After M. Mclaren e al.,“Advances in deep neural network approaches to speaker recognition” ICASSP 2015.

Denoising Autoencoder BN Features

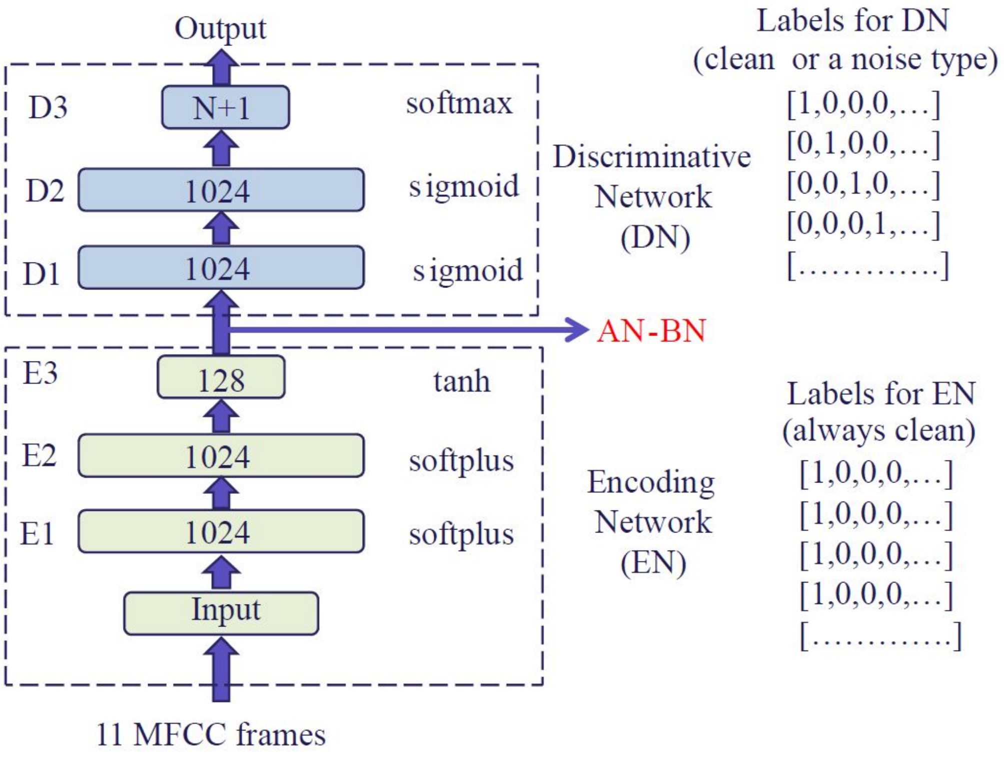
Denoising autoencoder for cepstral domain dereverberation.

- 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

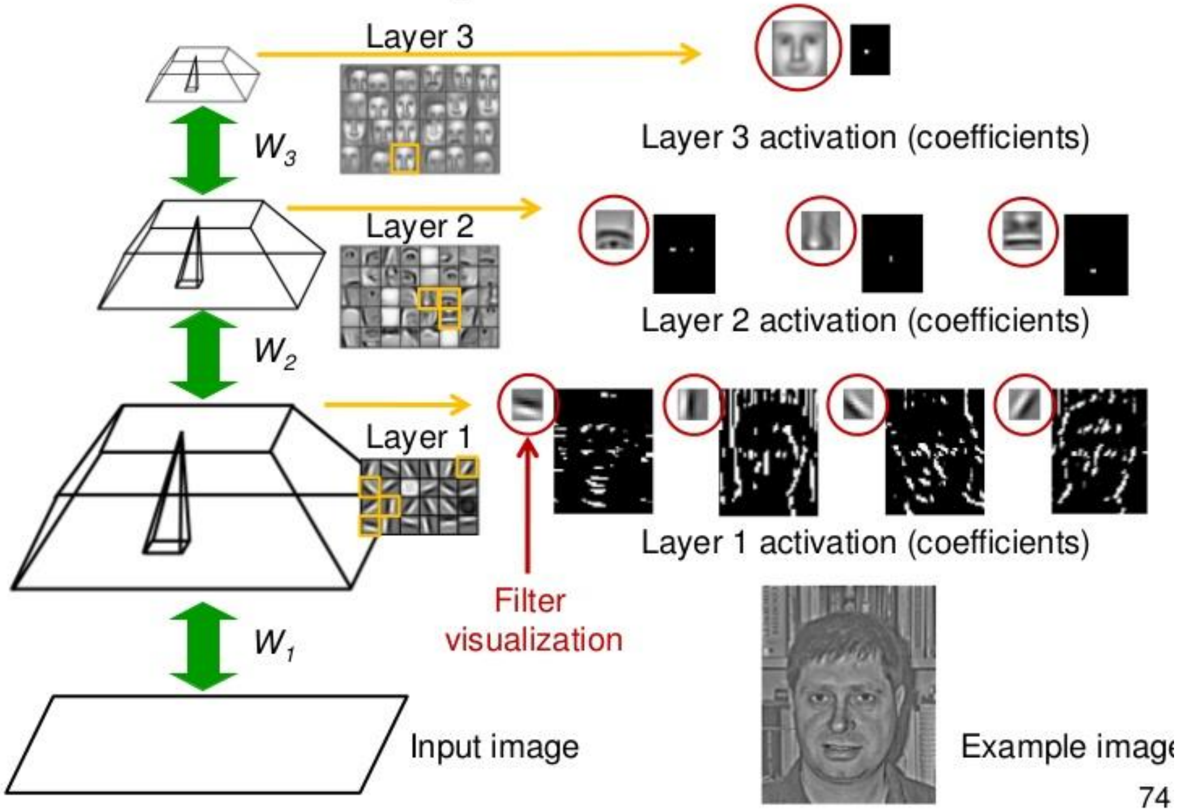


Adversarial Networks BN Features

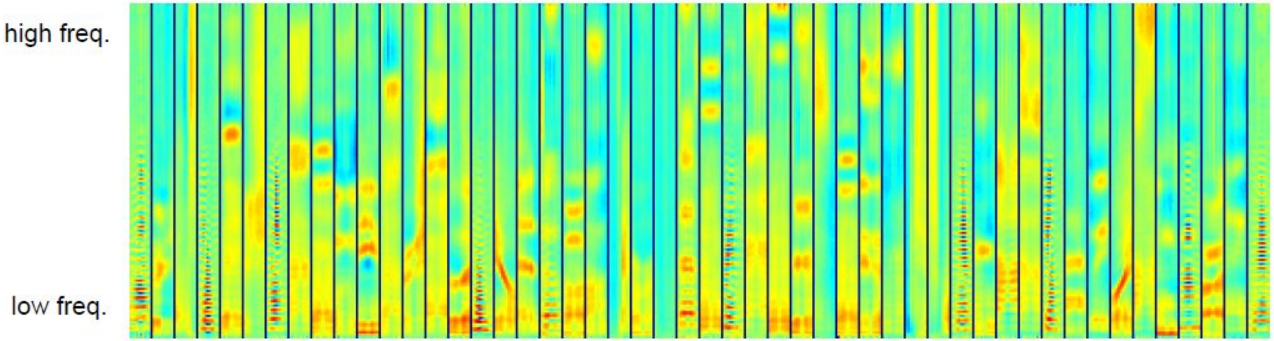


H. Yu, Z-H. Tan, Z. Ma, J. Guo, Adversarial Network Bottleneck Features for Noise Robust Speaker Verification, Proc. INTERSPEECH 2017

CDBN Features



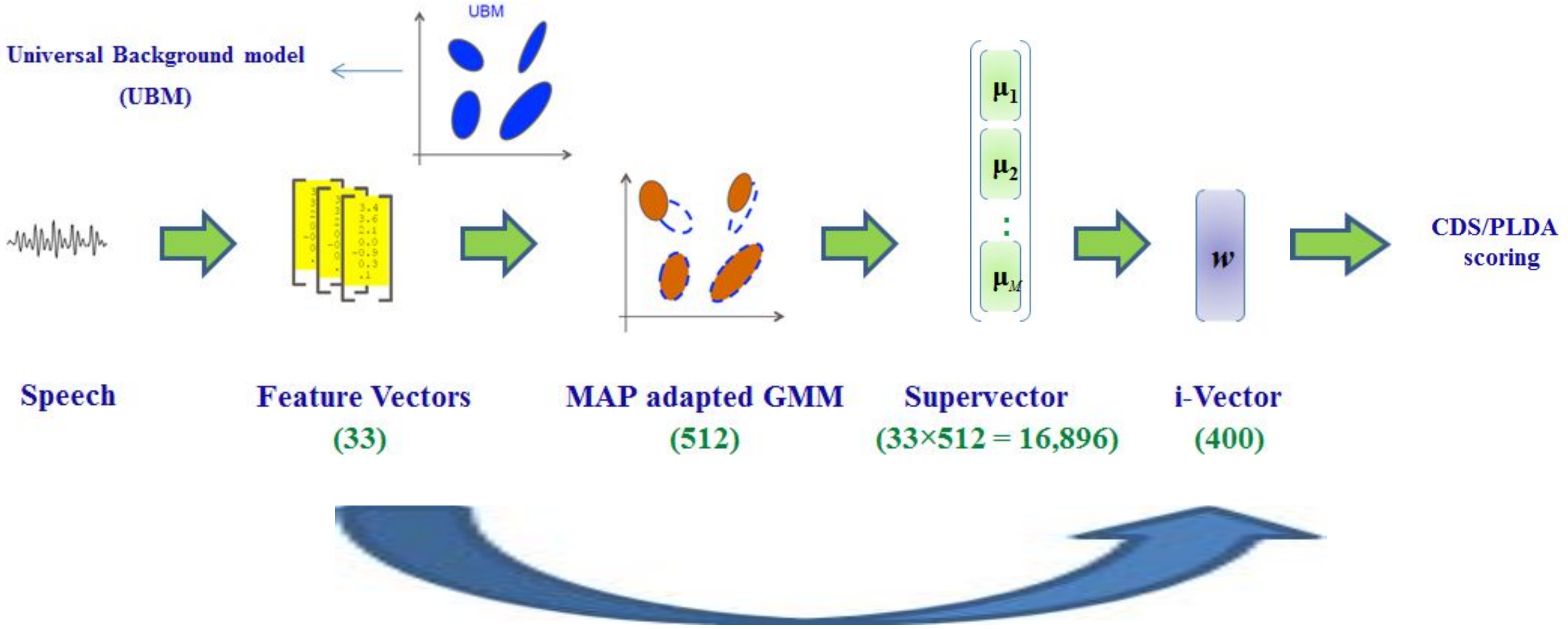
CDBN Features



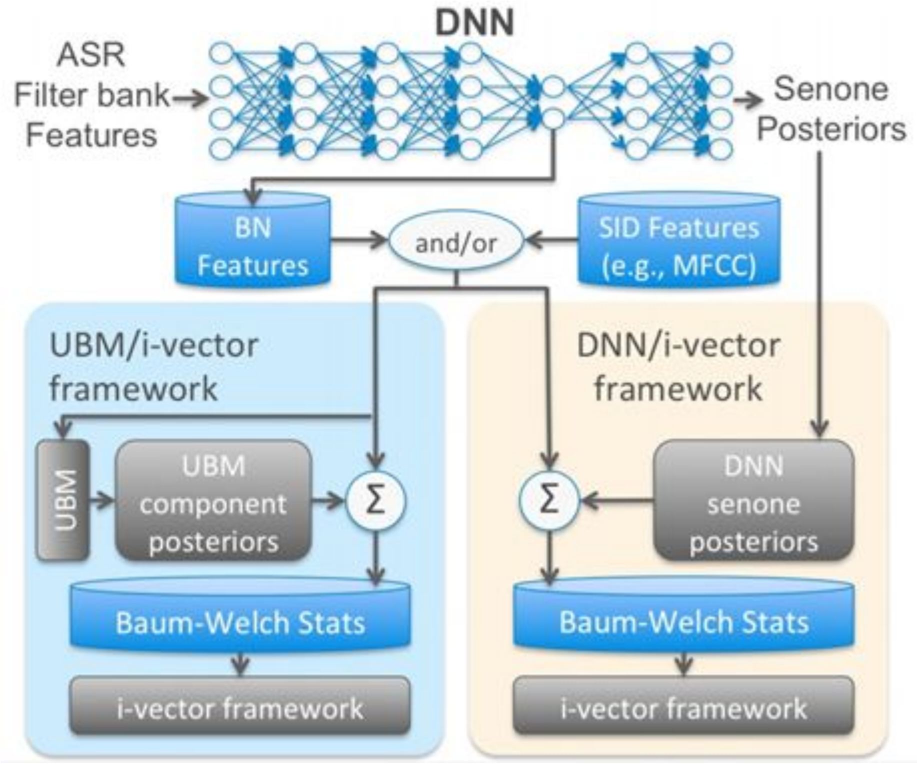
randomly selected first-layer CDBN bases

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

Front-End: i-vector Extraction

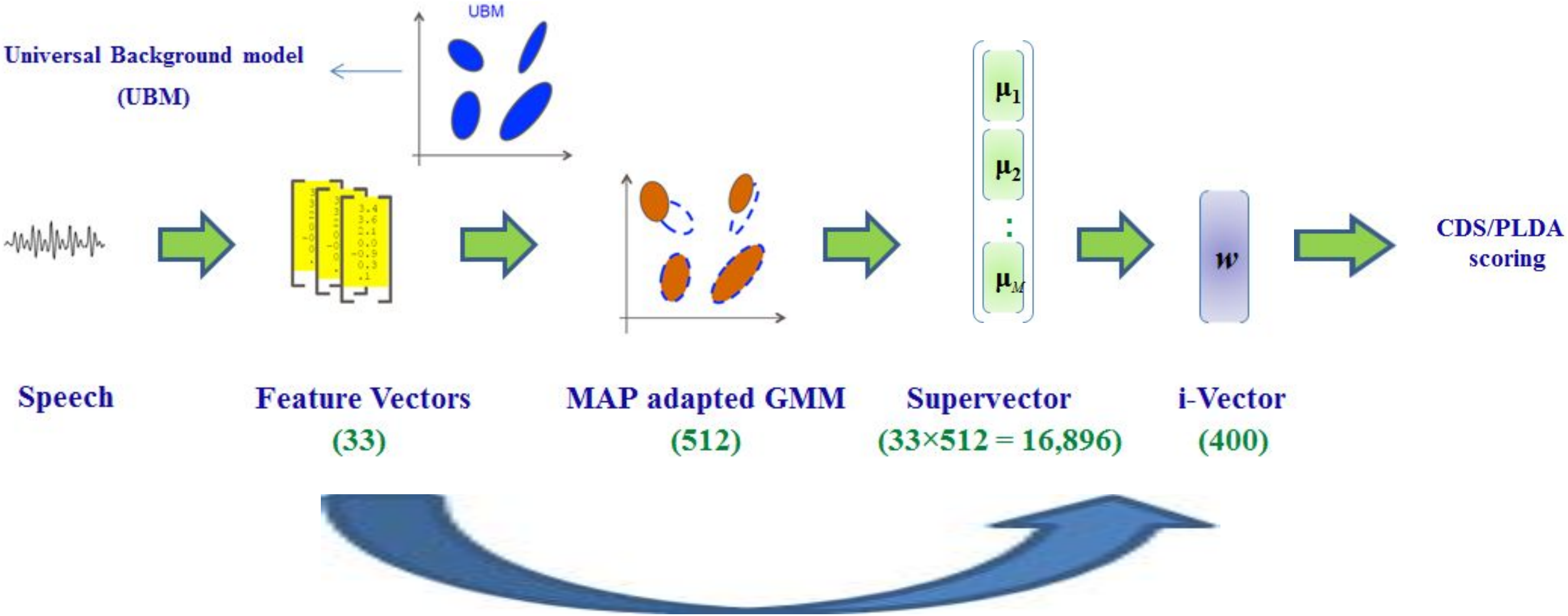


DL Front-End: i-vectors

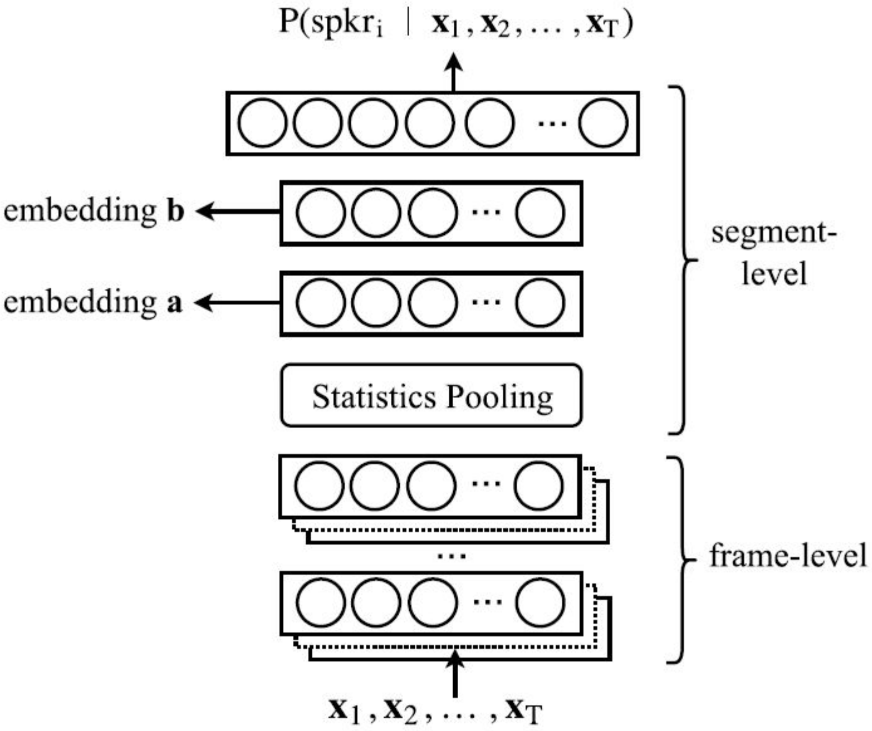


After M. Mclaren e al.,“Advances in deep neural network approaches to speaker recognition” ICASSP 2015.

Front-End: Features to Embeddings

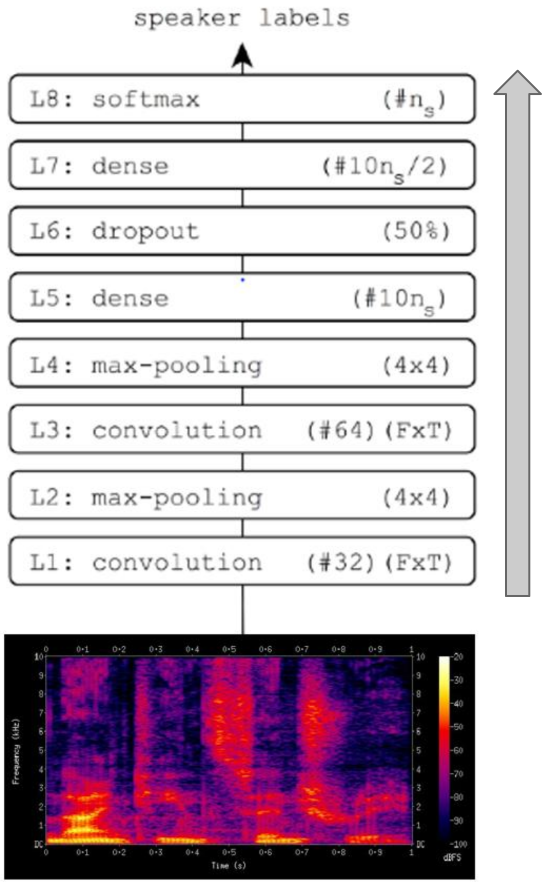


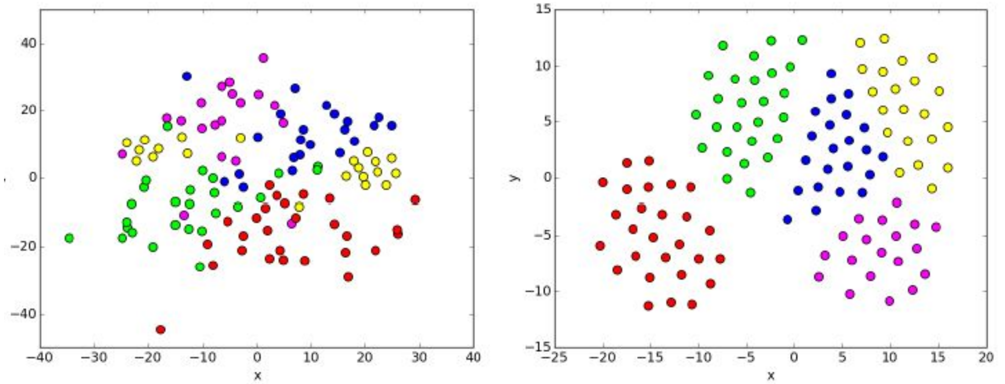
DNN Embeddings -> x-vectors



D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep Neural Network Embeddings for Text-Independent Speaker Verification, Proc. INTERSPEECH 2017

CNN Embeddings

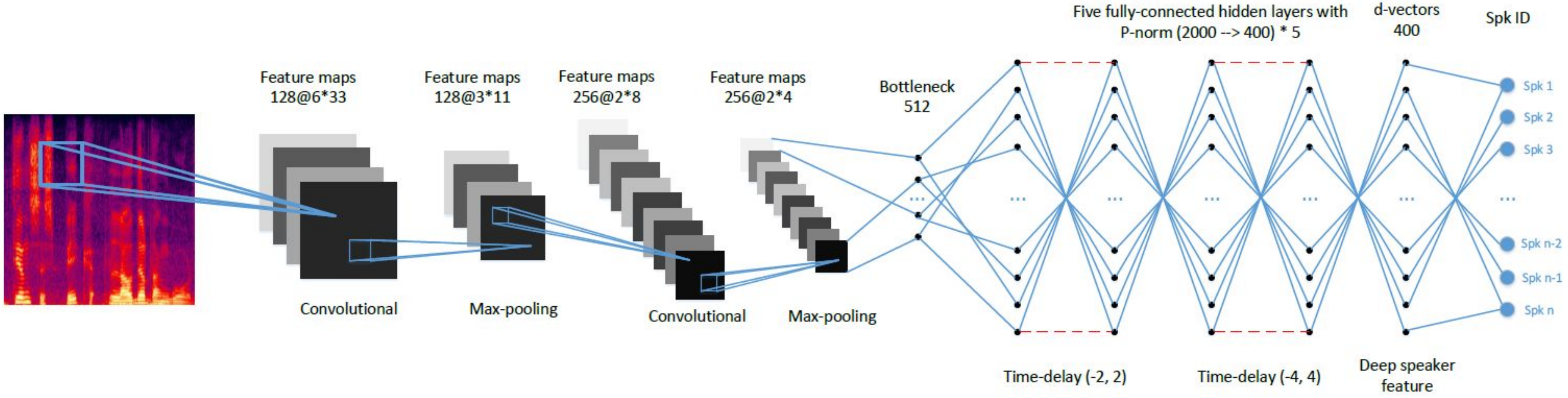




Five Speaker representations in 2 dimensions.
Left figure show the output vector of the softmax layer L8.
Right figure correspond to the same output vector of L5 dense layer.
Differents colors are assigned to different speakers.

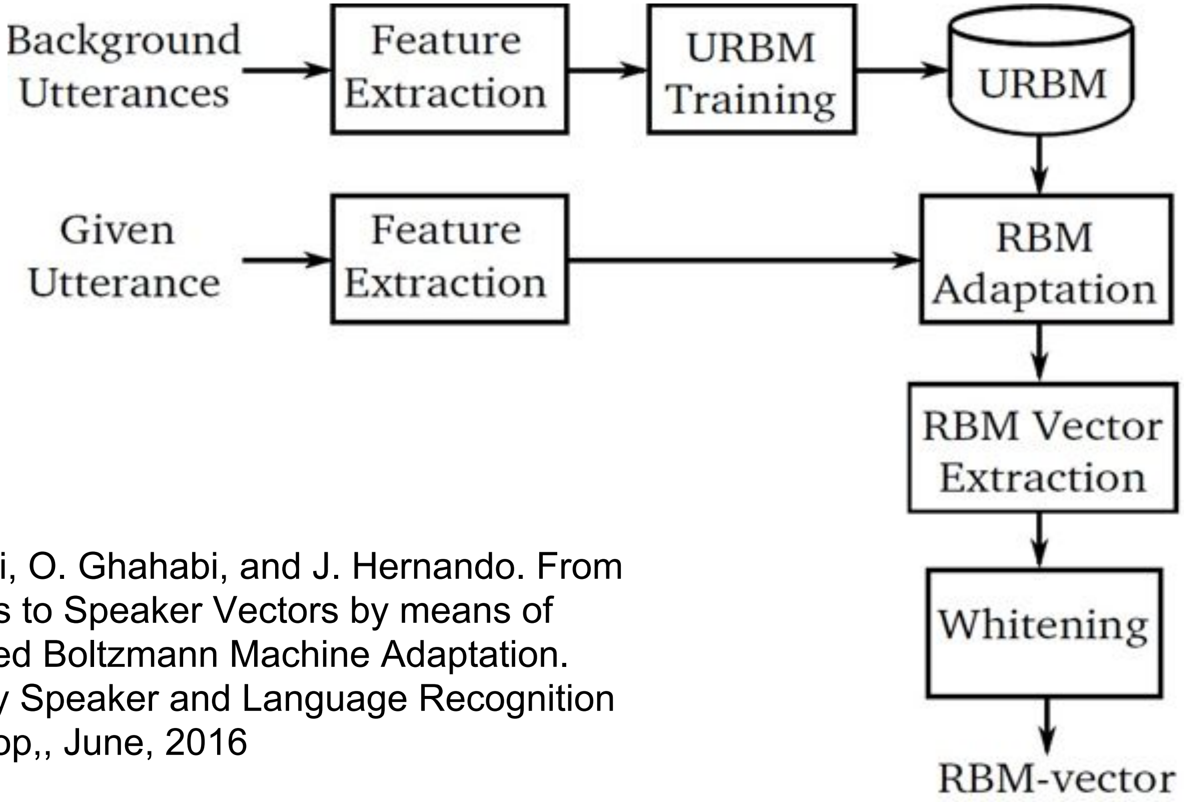
Yanik Lukic et al. “Speaker Identification and Clustering using Convolutional Neural Networks”. In 2016 IEEE International workshop on machine learning for signal processing. (2016)

Convolutional Time-Delay DNN Embeddings



L Li, Y Chen, Y Shi, Z Tang, D Wang, Deep Speaker Feature Learning for Text-independent Speaker Verification, Proc. INTERSPEECH 2017

RBM Embeddings



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

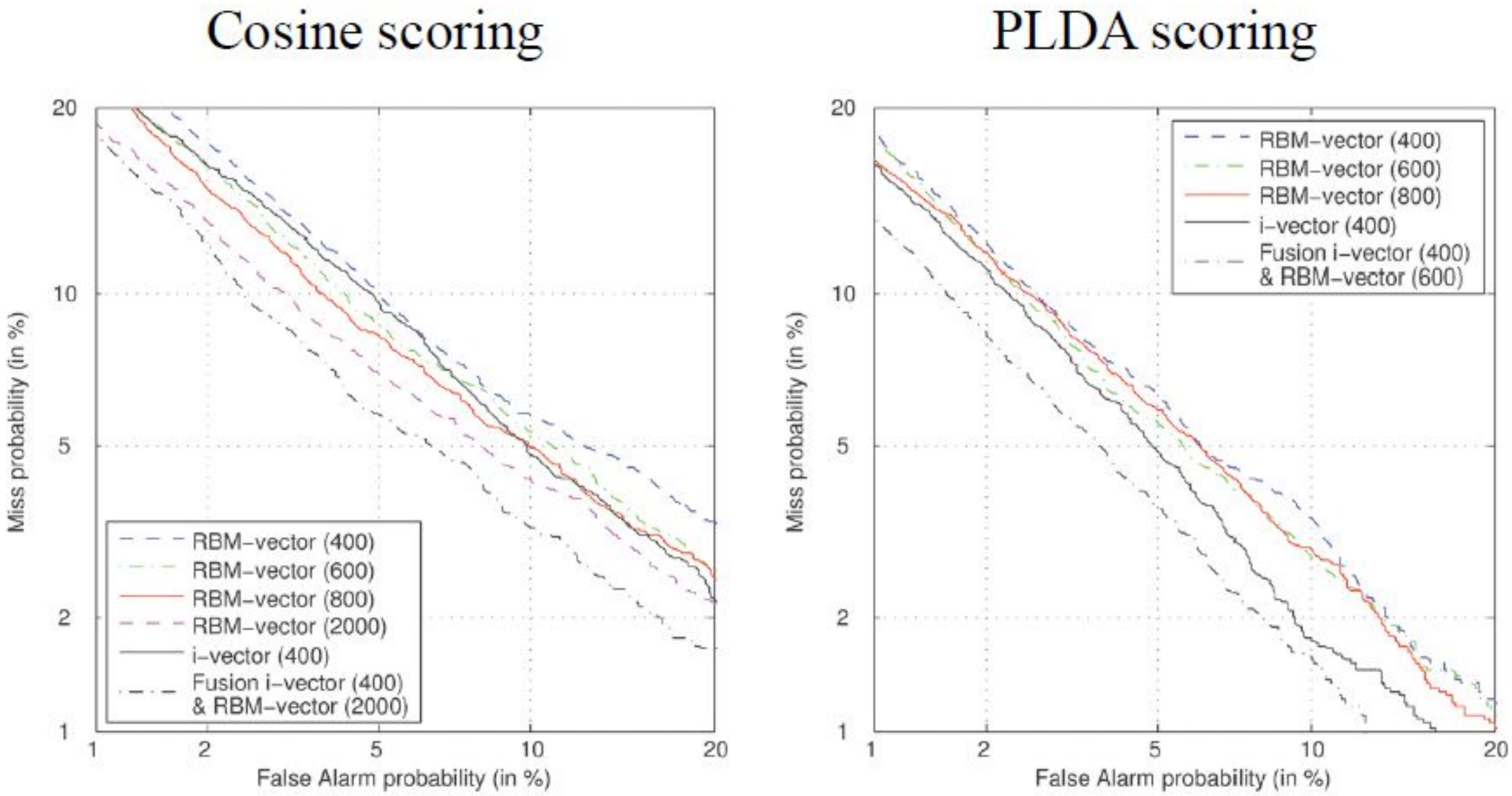
RBM Embeddings

- RBM supervectors

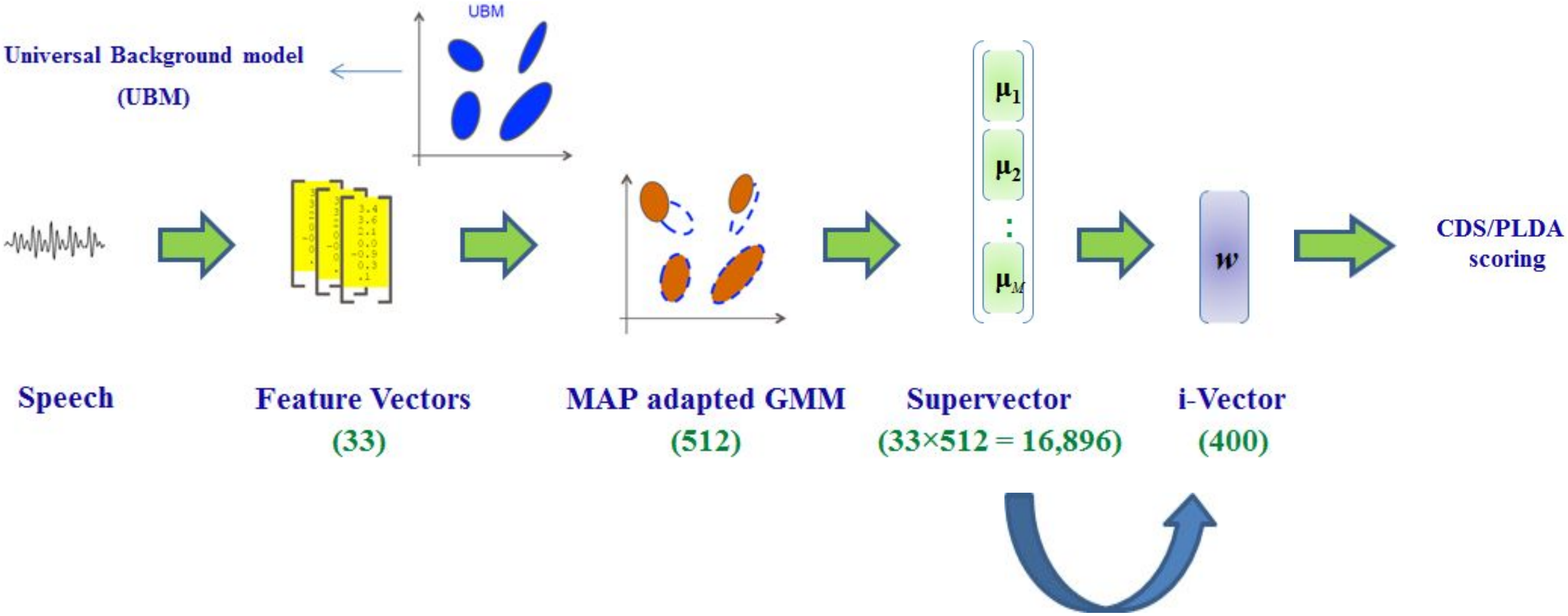
$$\begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1h} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{v1} & \omega_{v2} & \cdots & \omega_{vh} \end{pmatrix} \qquad \begin{pmatrix} \begin{pmatrix} \omega_{11} \\ \omega_{12} \\ \vdots \\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

- 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

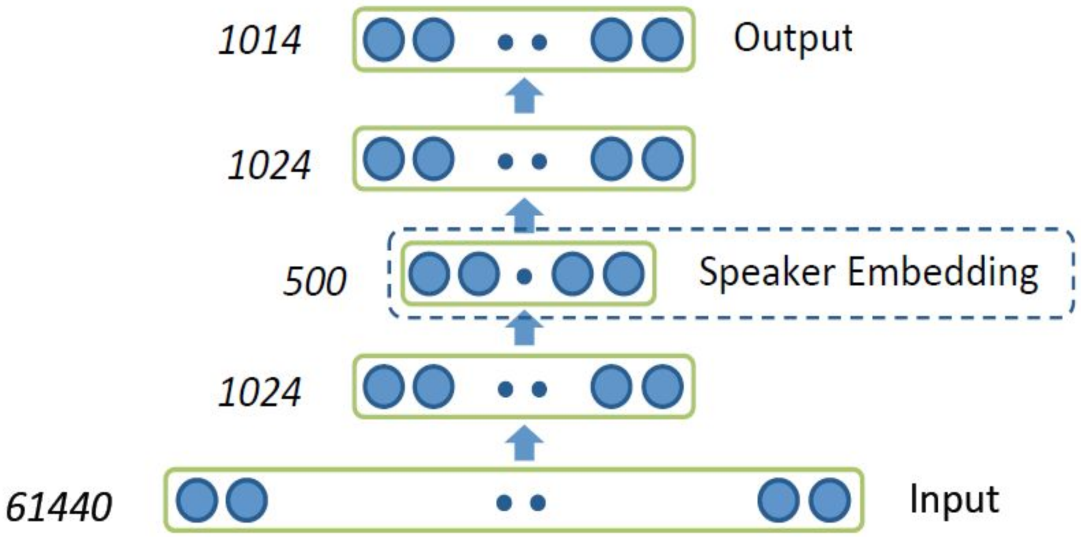
RBM Embeddings



Front-End: Vectors to Embeddings



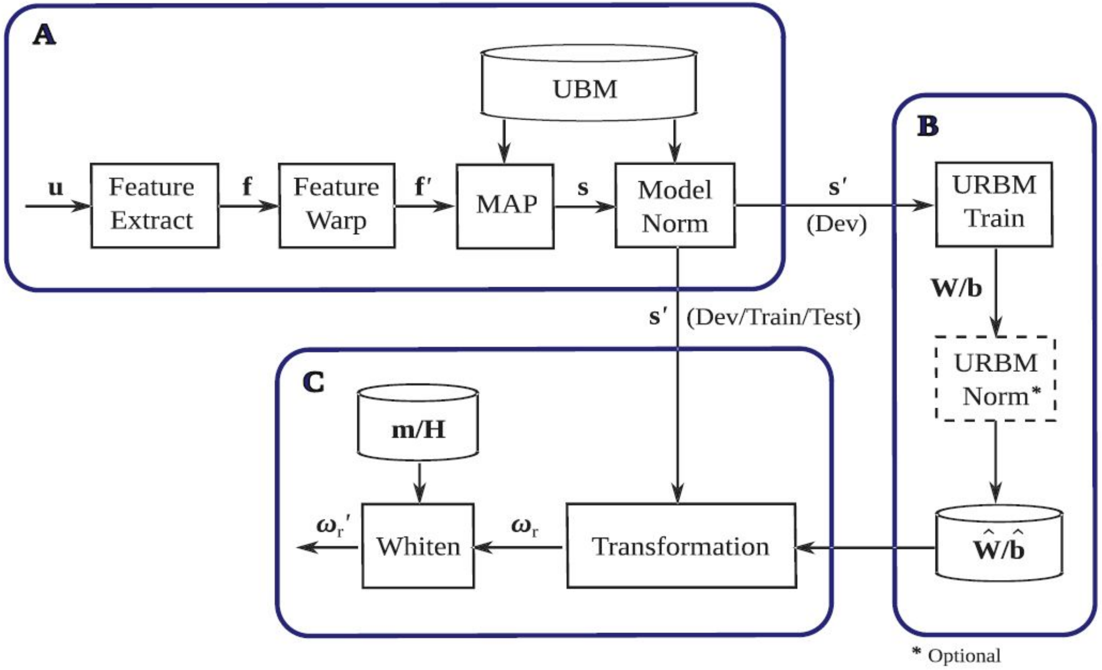
From Supervectors to Embeddings



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

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

GMM-RBM vectors



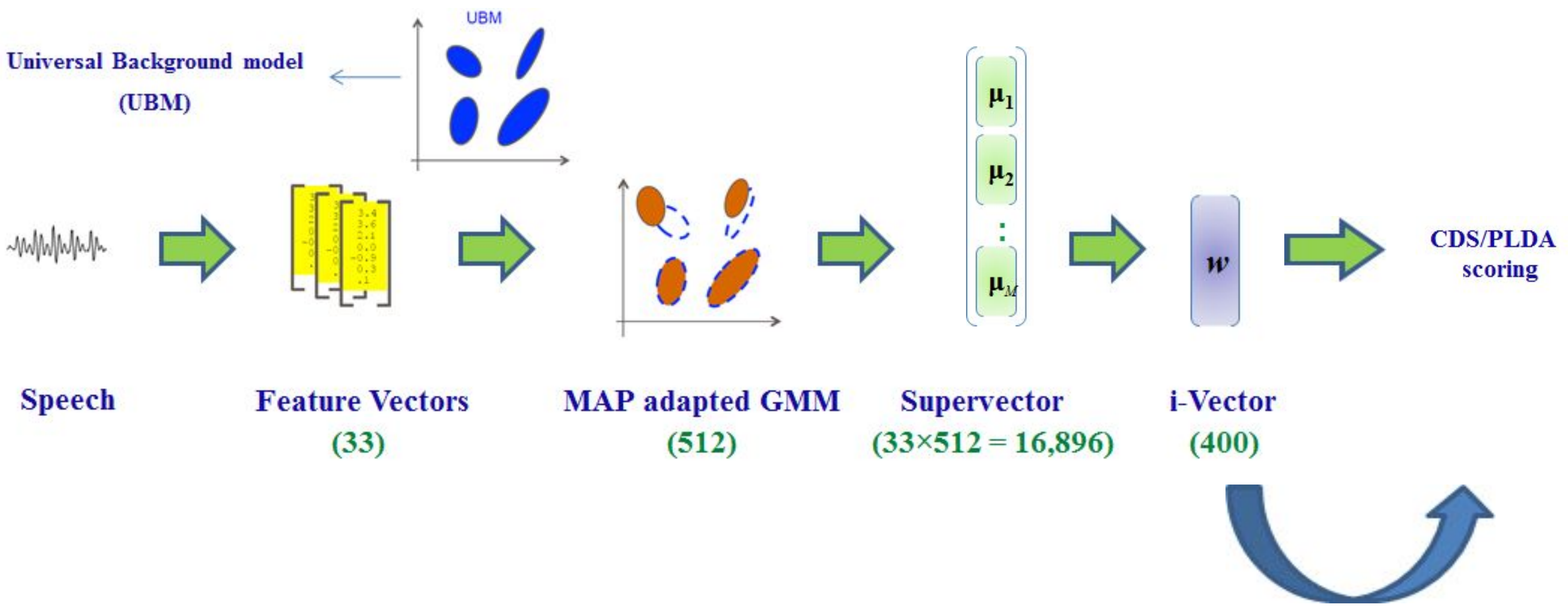
O. Ghahabi, J. Hernando, Restricted Boltzmann machines for vector representation of speech in speaker recognition, Computer Speech & Language, 47 (2018) 16-29

GMM-RBM vectors

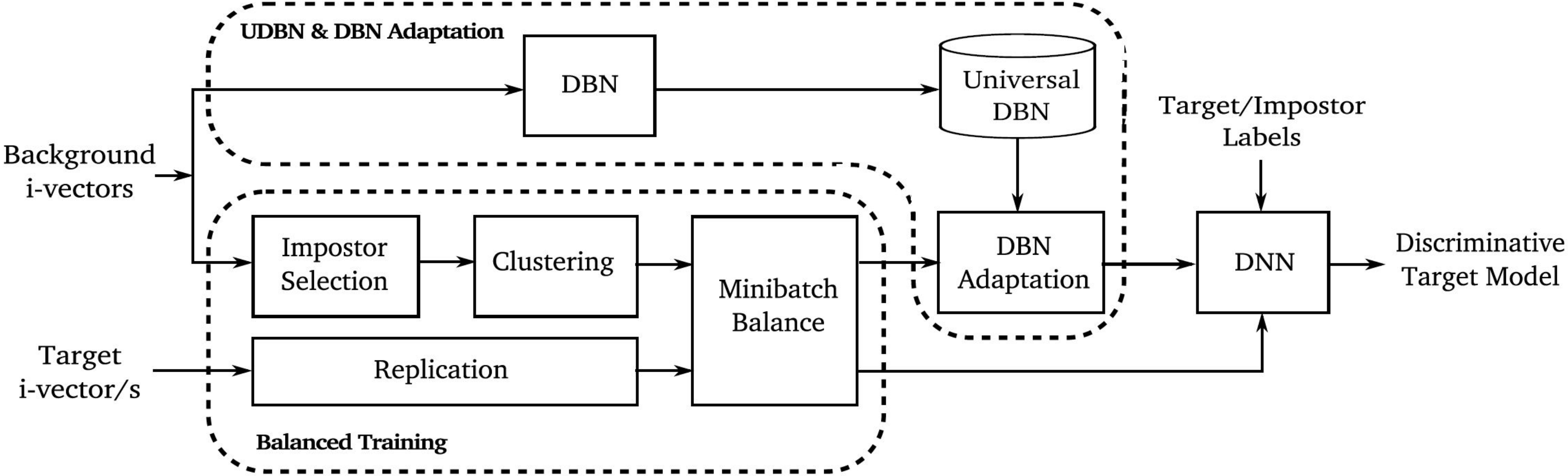
Performance comparison of proposed GMM–RBM vectors and conventional i-vectors on the **evaluation** set core test condition-common 5 of NIST 2010 SRE. GMM–RBM vectors and i-vectors are of a same size of 400.

		Cosine		PLDA	
		EER (%)	minDCF	EER (%)	minDCF
[1]	i-Vector	6.270	0.05450	4.096	0.04993
[2]	GMM–RBM vector (trained with ReLU)	6.638	0.06228	4.517	0.05085
[3]	GMM–RBM vector (trained with VReLU)	6.497	0.06099	3.907	0.05184
Fusion [1] and [3]		**5.791**	**0.05238**	**3.814**	**0.04673**

Back-End

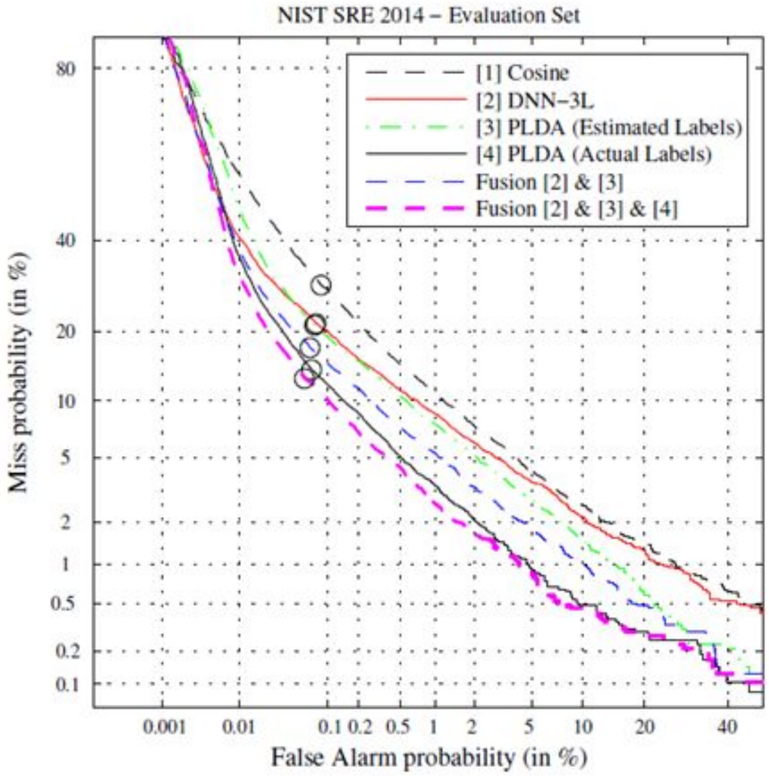

UBM
Universal Background model
(UBM)
μ_1
μ_2
μ_M
w
CDS/PLDA scoring
Speech
Feature Vectors
(33)
MAP adapted GMM
(512)
Supervector
(33×512 = 16,896)
i-Vector
(400)

DNN i-Vector Back-End



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

DNN i-Vector Back-End



	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)