Speech Features per i task di Speaker Identification e Verification

Autore: Gabriele Nicolò Costa

Outline

1

Introduzione al problema

2

Speech features e speech embeddings

3

Speaker Models e Backend – Architettura proposta

4

Risultati sperimentali

Speaker Identification

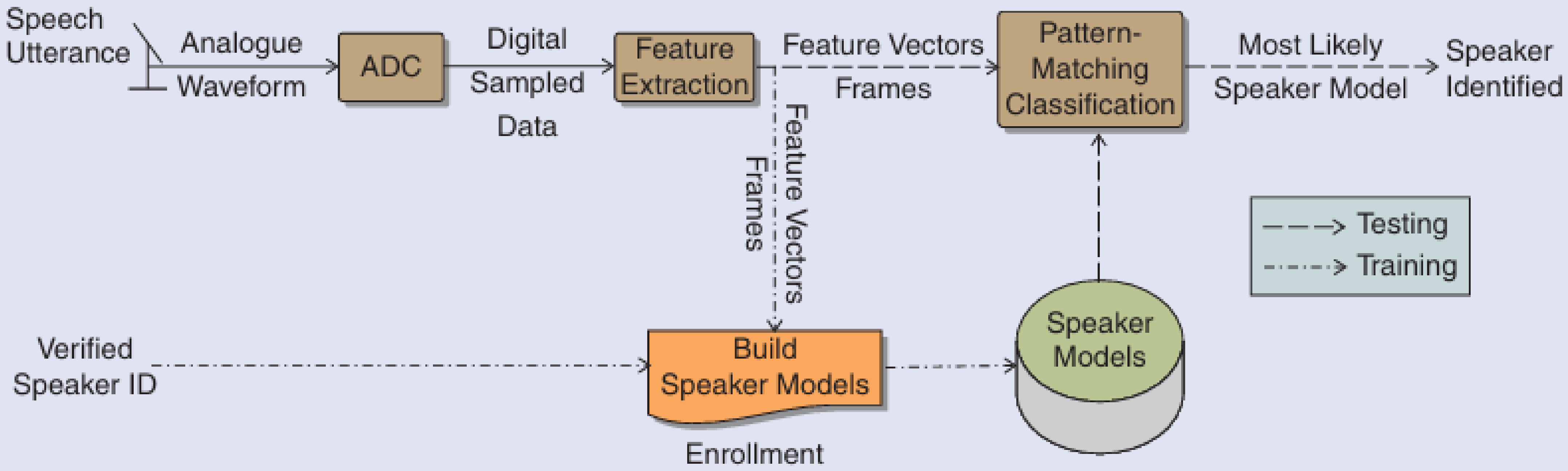


Figure 1. Block Diagram of a Speaker Identification System.

Speaker Verification

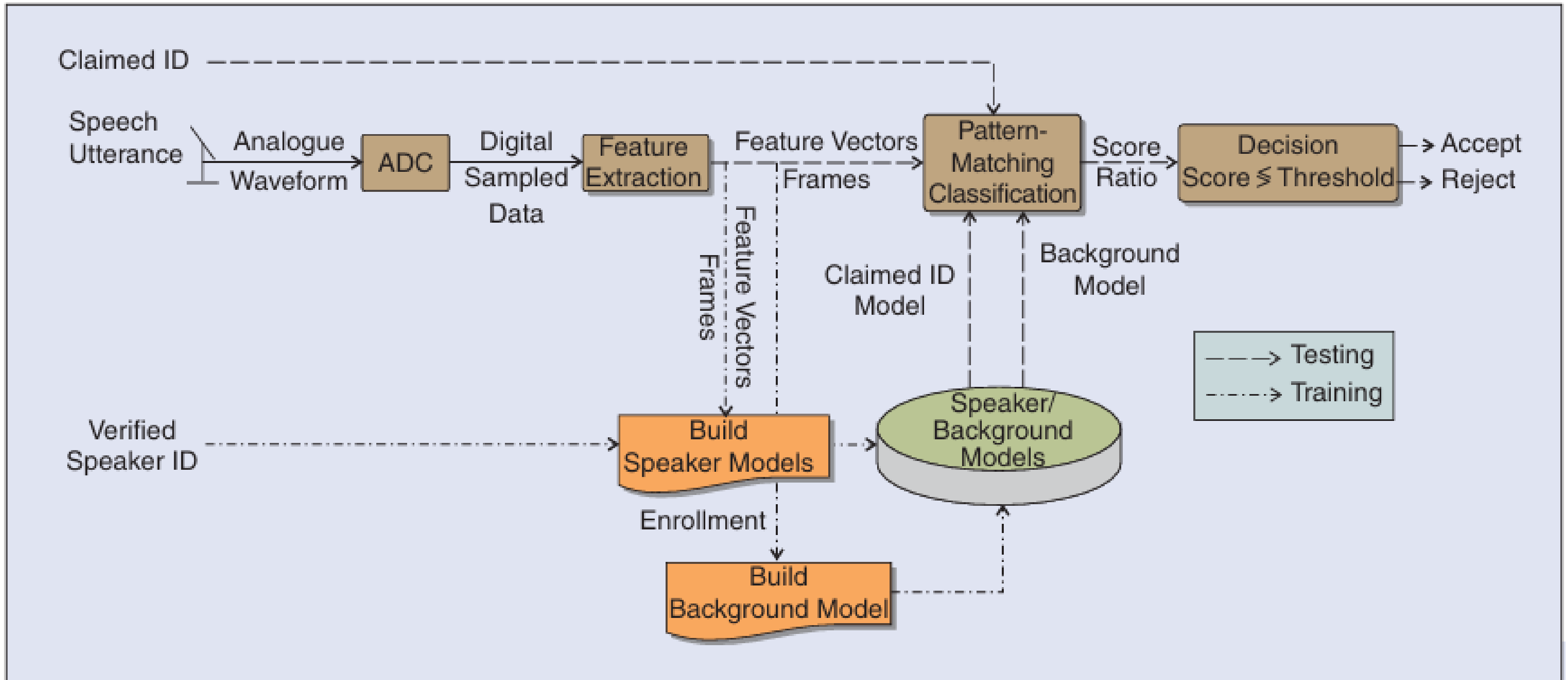
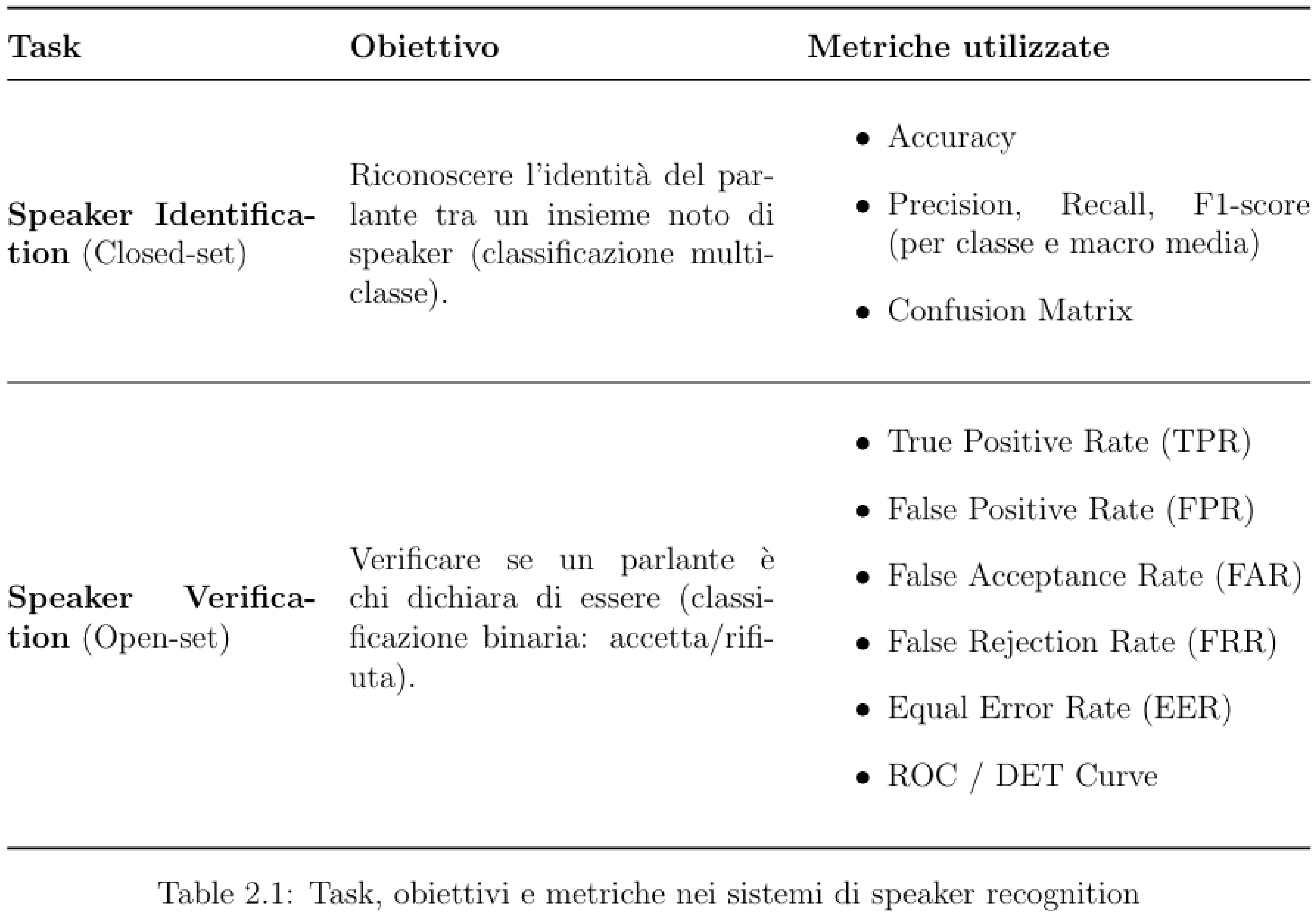


Figure 2. Block diagram of a speaker verification system.

Metriche e Valutazione dei sistemi



Task	**Obiettivo**	**Metriche utilizzate**
Speaker Identification (Closed-set)	Riconoscere l'identità del parlante tra un insieme noto di speaker (classificazione multiclasse).	• Accuracy • Precision, Recall, F1-score (per classe e macro media) • Confusion Matrix
Speaker Verification (Open-set)	Verificare se un parlante è chi dichiara di essere (classificazione binaria: accetta/rifiuta).	• True Positive Rate (TPR) • False Positive Rate (FPR) • False Acceptance Rate (FAR) • False Rejection Rate (FRR) • Equal Error Rate (EER) • ROC / DET Curve

Table 2.1: Task, obiettivi e metriche nei sistemi di speaker recognition

Speech Features

- Cepstrum
- FilterBanks e Mel-scale
- Mel-Frequency Cepstral Coefficients (MFCCs)

Framework estrazione Features

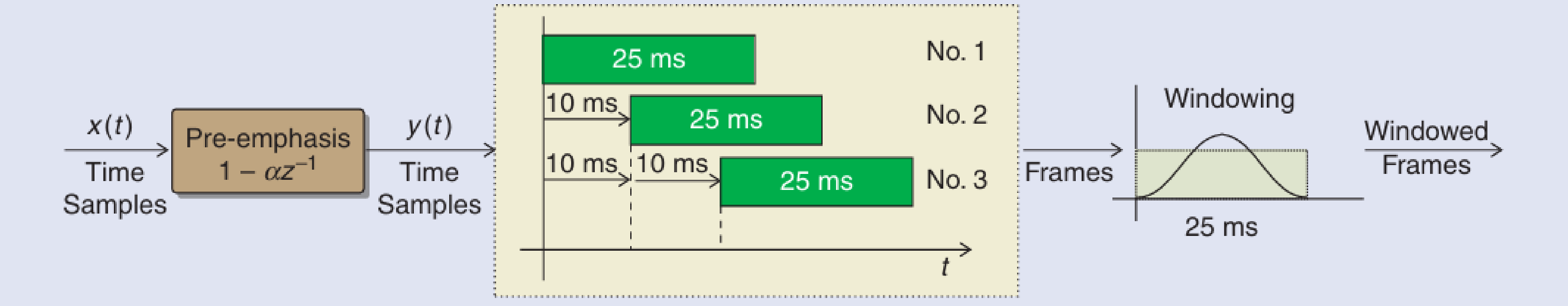


Figure 3. Analysis block diagram for framing.

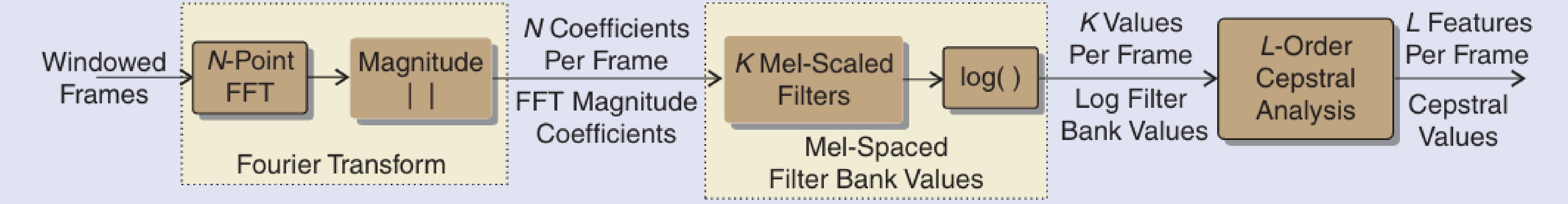


Figure 4. Analysis Block Diagram for MFCC feature vectors.

Speech Embeddings

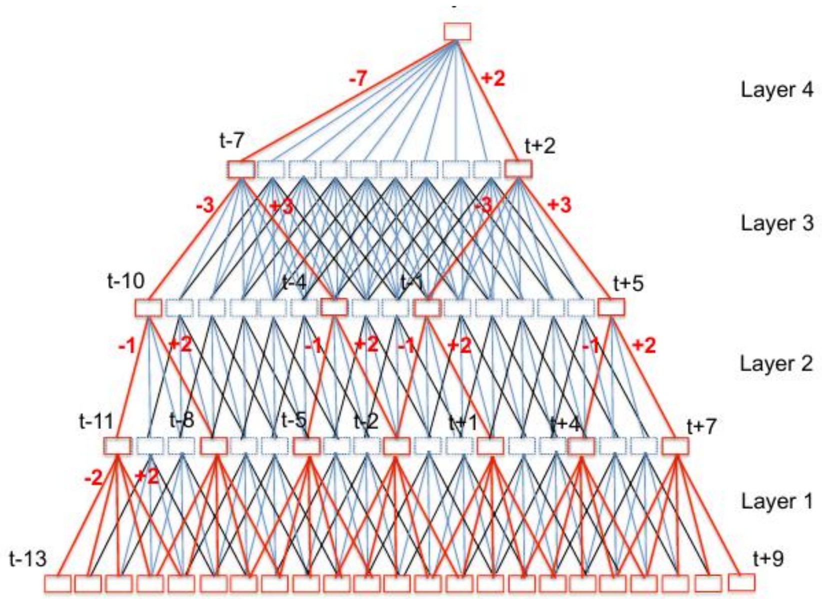


Figure 1: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red)

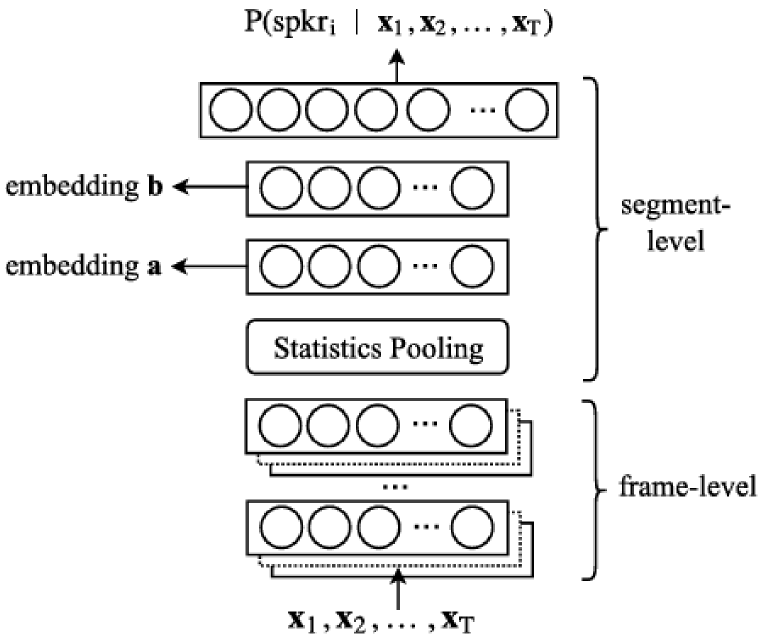

$\mathrm{P}(\mathrm{spkr}_i \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$
embedding **b**
embedding **a**
Statistics Pooling
segment-level
frame-level
$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$

Figure 1: *Diagram of the DNN. Segment-level embeddings (e.g.,* $\mathbf{a}$ *or* $\mathbf{b}$*) can be extracted from any layer of the network after the statistics pooling layer.*

Speaker Identificatione Verification pipieline

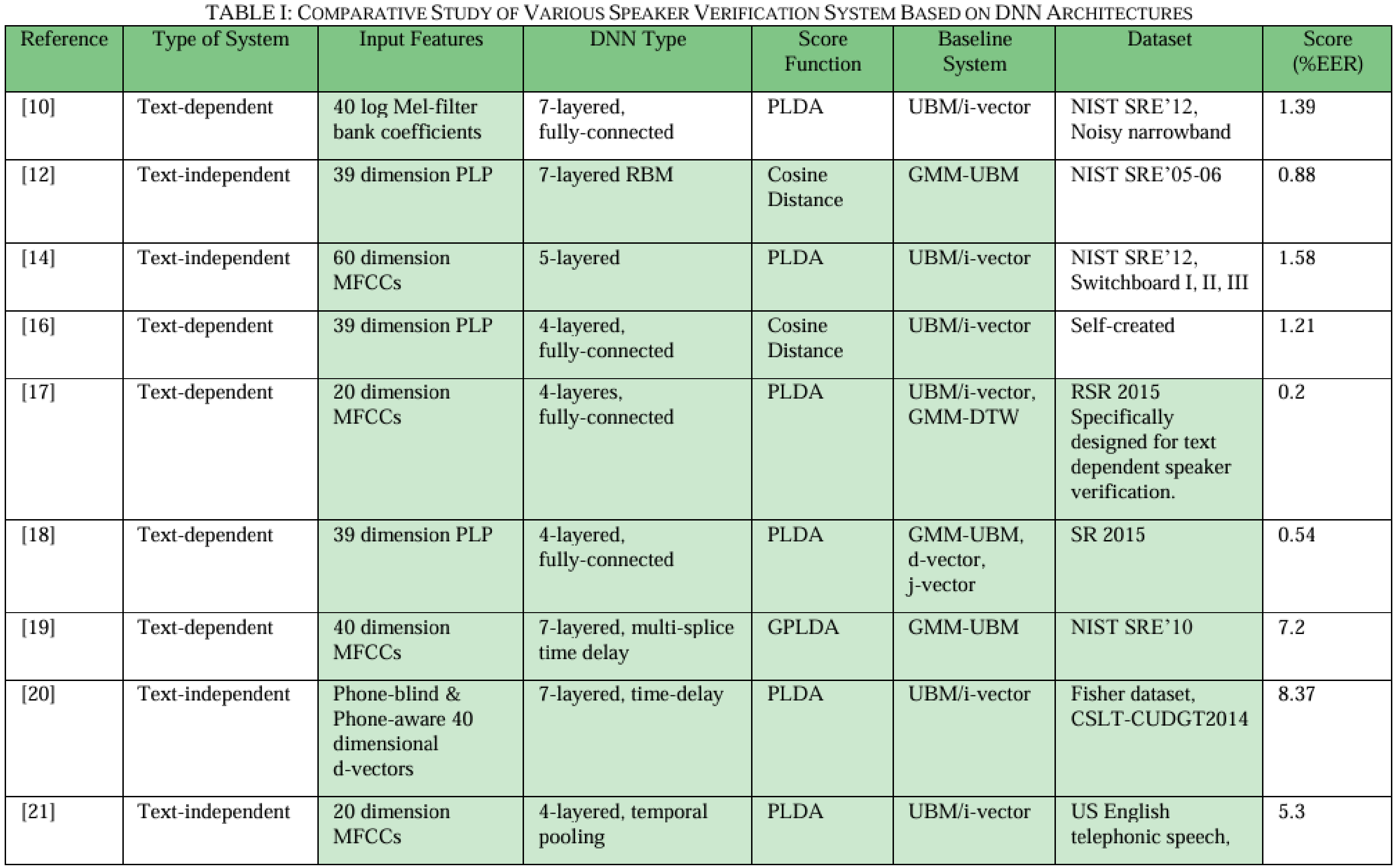


TABLE I: COMPARATIVE STUDY OF VARIOUS SPEAKER VERIFICATION SYSTEM BASED ON DNN ARCHITECTURES

Reference	Type of System	Input Features	DNN Type	Score Function	Baseline System	Dataset	Score (%EER)
[10]	Text-dependent	40 log Mel-filter bank coefficients	7-layered, fully-connected	PLDA	UBM/i-vector	NIST SRE'12, Noisy narrowband	1.39
[12]	Text-independent	39 dimension PLP	7-layered RBM	Cosine Distance	GMM-UBM	NIST SRE'05-06	0.88
[14]	Text-independent	60 dimension MFCCs	5-layered	PLDA	UBM/i-vector	NIST SRE'12, Switchboard I, II, III	1.58
[16]	Text-dependent	39 dimension PLP	4-layered, fully-connected	Cosine Distance	UBM/i-vector	Self-created	1.21
[17]	Text-dependent	20 dimension MFCCs	4-layeres, fully-connected	PLDA	UBM/i-vector, GMM-DTW	RSR 2015 Specifically designed for text dependent speaker verification.	0.2
[18]	Text-dependent	39 dimension PLP	4-layered, fully-connected	PLDA	GMM-UBM, d-vector, j-vector	SR 2015	0.54
[19]	Text-dependent	40 dimension MFCCs	7-layered, multi-splice time delay	GPLDA	GMM-UBM	NIST SRE'10	7.2
[20]	Text-independent	Phone-blind & Phone-aware 40 dimensional d-vectors	7-layered, time-delay	PLDA	UBM/i-vector	Fisher dataset, CSLT-CUDGT2014	8.37
[21]	Text-independent	20 dimension MFCCs	4-layered, temporal pooling	PLDA	UBM/i-vector	US English telephonic speech,	5.3

Speaker Models e Background models

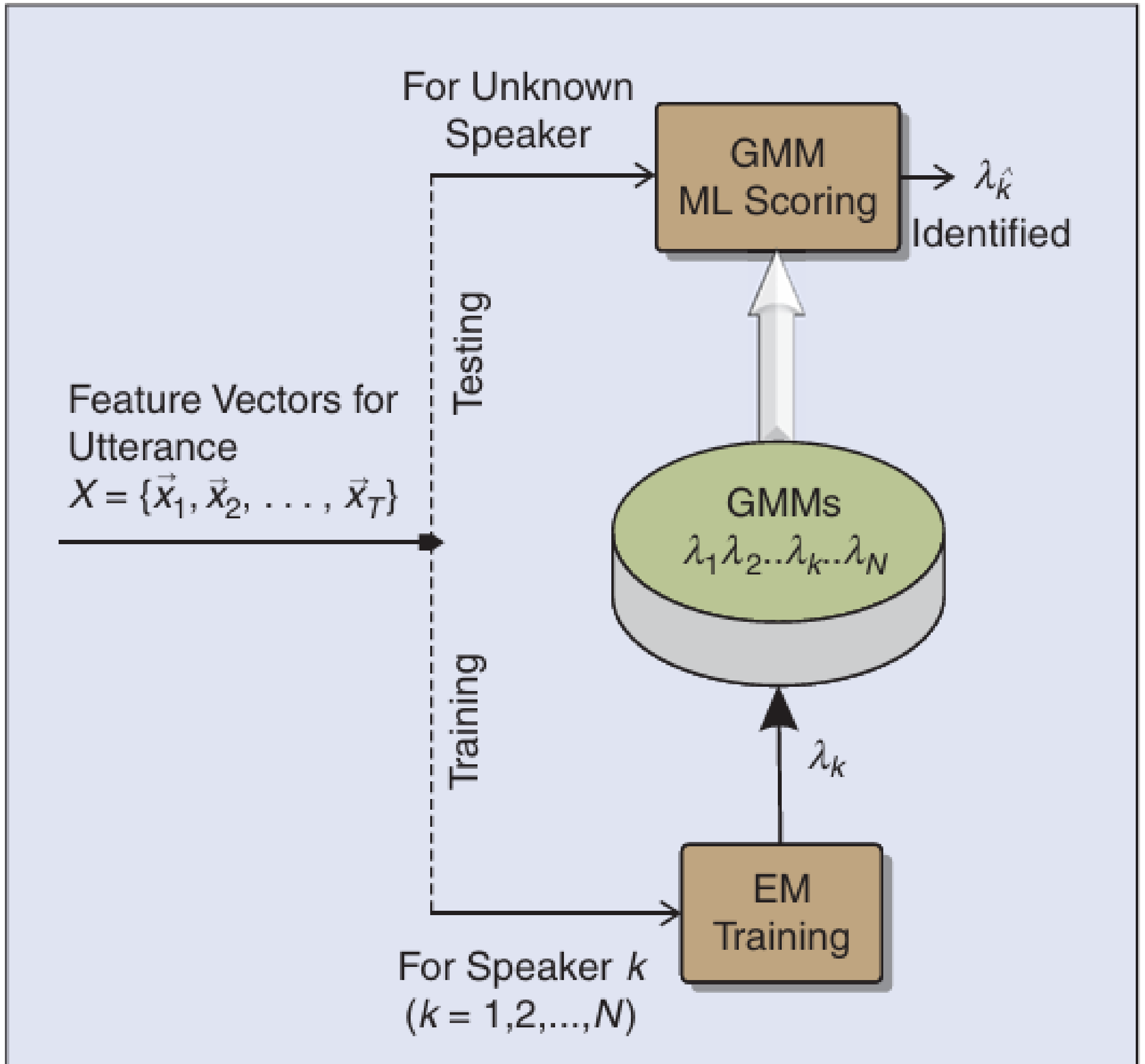

For Unknown Speaker
GMM ML Scoring
$\lambda_{\hat{k}}$
Identified
Testing
Feature Vectors for Utterance $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$
GMMs
$\lambda_1 \lambda_2 .. \lambda_k .. \lambda_N$
Training
λ_k
EM Training
For Speaker k ($k = 1,2,\ldots,N$)

Figure 8. GMM training and testing framework assuming *N* speakers.

Architettura proposta

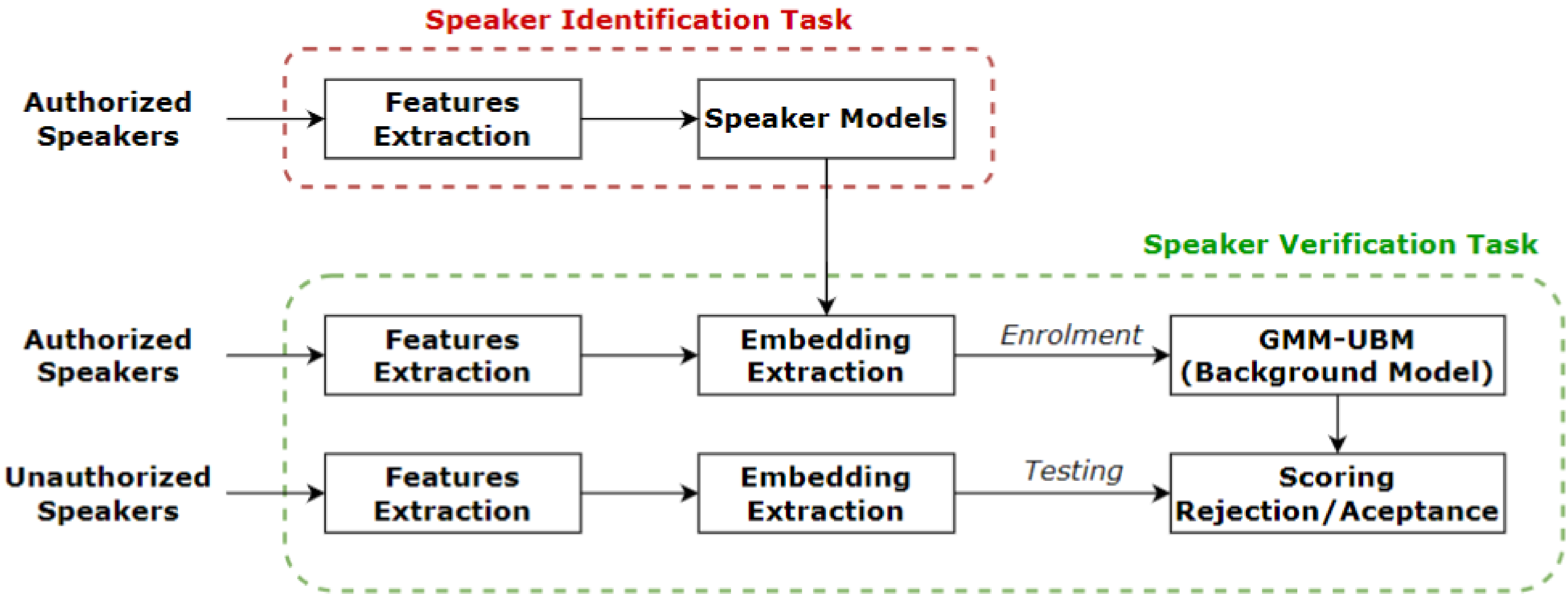
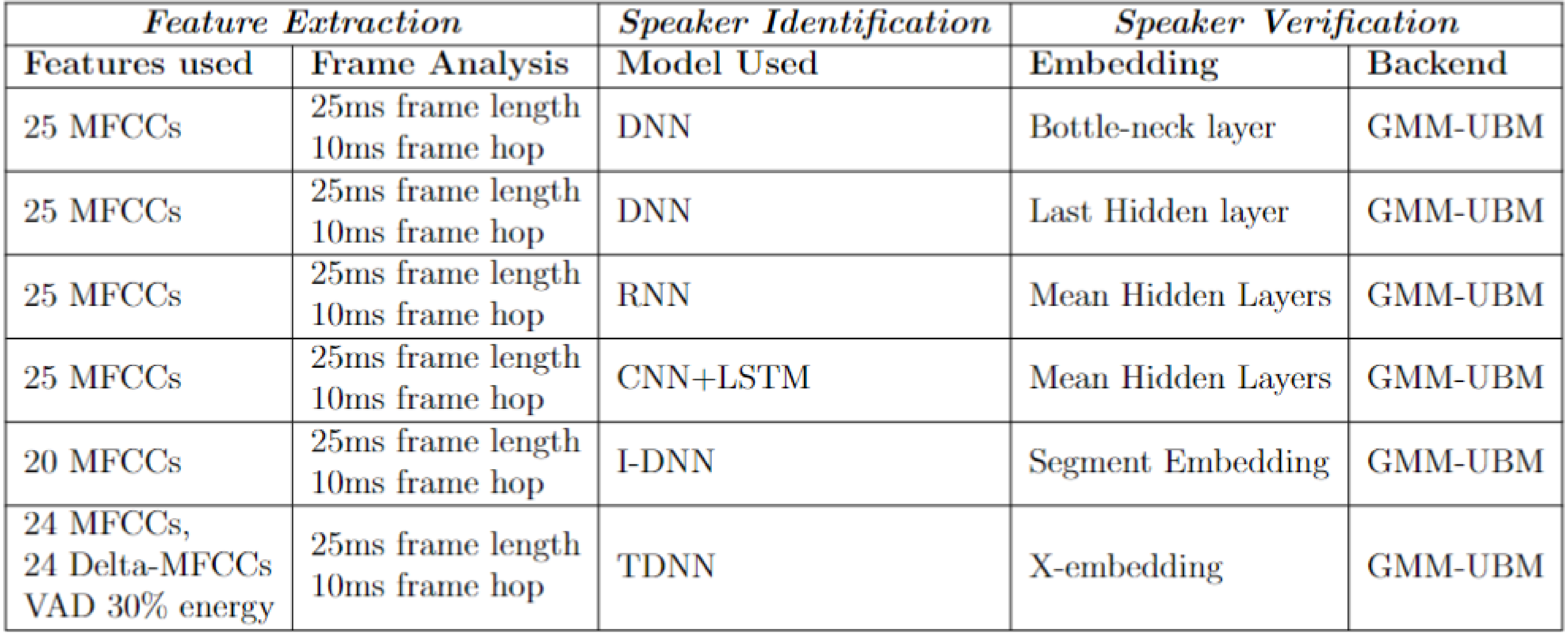


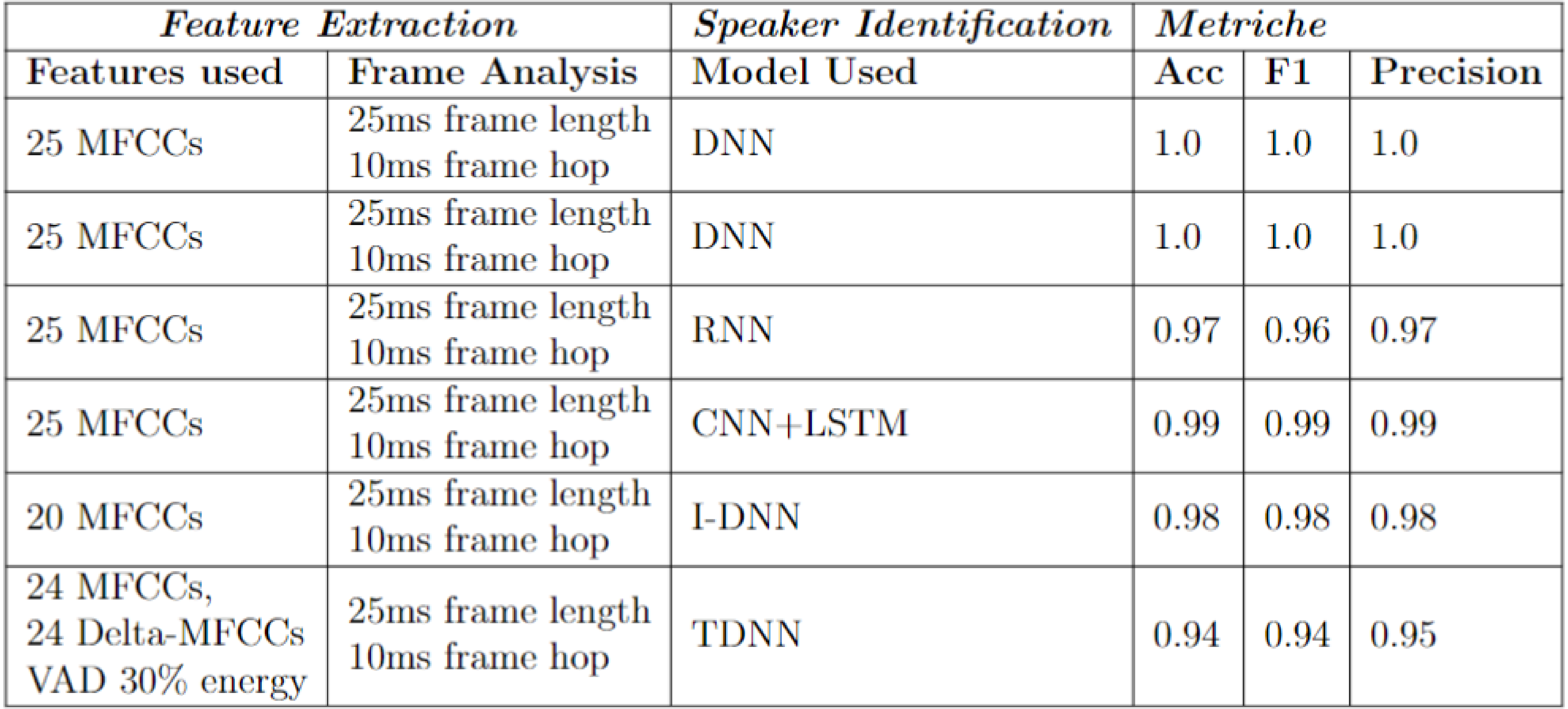
Figure 4.2: Architettura del progetto proposta

Risultati sperimentali



Feature Extraction		*Speaker Identification*	*Speaker Verification*	
Features used	**Frame Analysis**	**Model Used**	**Embedding**	**Backend**
25 MFCCs	25ms frame length 10ms frame hop	DNN	Bottle-neck layer	GMM-UBM
25 MFCCs	25ms frame length 10ms frame hop	DNN	Last Hidden layer	GMM-UBM
25 MFCCs	25ms frame length 10ms frame hop	RNN	Mean Hidden Layers	GMM-UBM
25 MFCCs	25ms frame length 10ms frame hop	CNN+LSTM	Mean Hidden Layers	GMM-UBM
20 MFCCs	25ms frame length 10ms frame hop	I-DNN	Segment Embedding	GMM-UBM
24 MFCCs, 24 Delta-MFCCs VAD 30% energy	25ms frame length 10ms frame hop	TDNN	X-embedding	GMM-UBM

Table 5.1: Riassunto degli esperimenti condotti



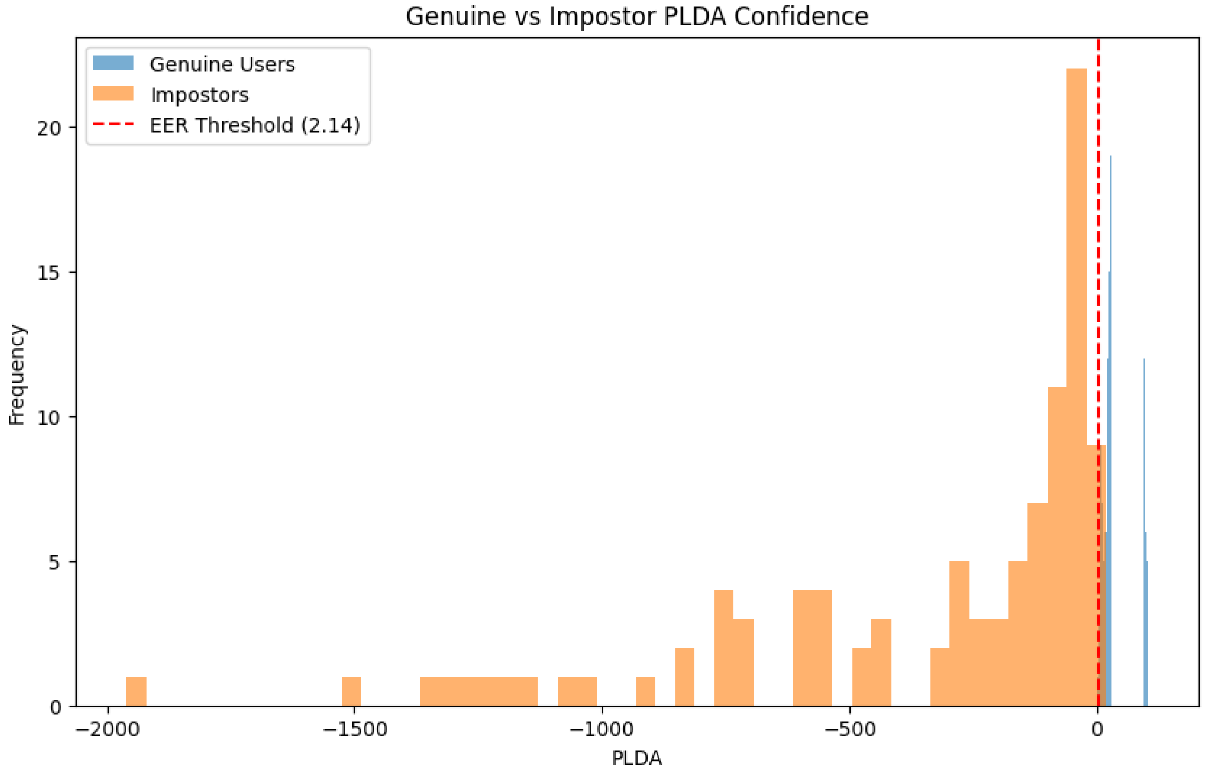
Feature Extraction		*Speaker Identification*	*Metriche*		
Features used	**Frame Analysis**	**Model Used**	**Acc**	**F1**	**Precision**
25 MFCCs	25ms frame length 10ms frame hop	DNN	1.0	1.0	1.0
25 MFCCs	25ms frame length 10ms frame hop	DNN	1.0	1.0	1.0
25 MFCCs	25ms frame length 10ms frame hop	RNN	0.97	0.96	0.97
25 MFCCs	25ms frame length 10ms frame hop	CNN+LSTM	0.99	0.99	0.99
20 MFCCs	25ms frame length 10ms frame hop	I-DNN	0.98	0.98	0.98
24 MFCCs, 24 Delta-MFCCs VAD 30% energy	25ms frame length 10ms frame hop	TDNN	0.94	0.94	0.95

Table 5.2: Speaker Identification

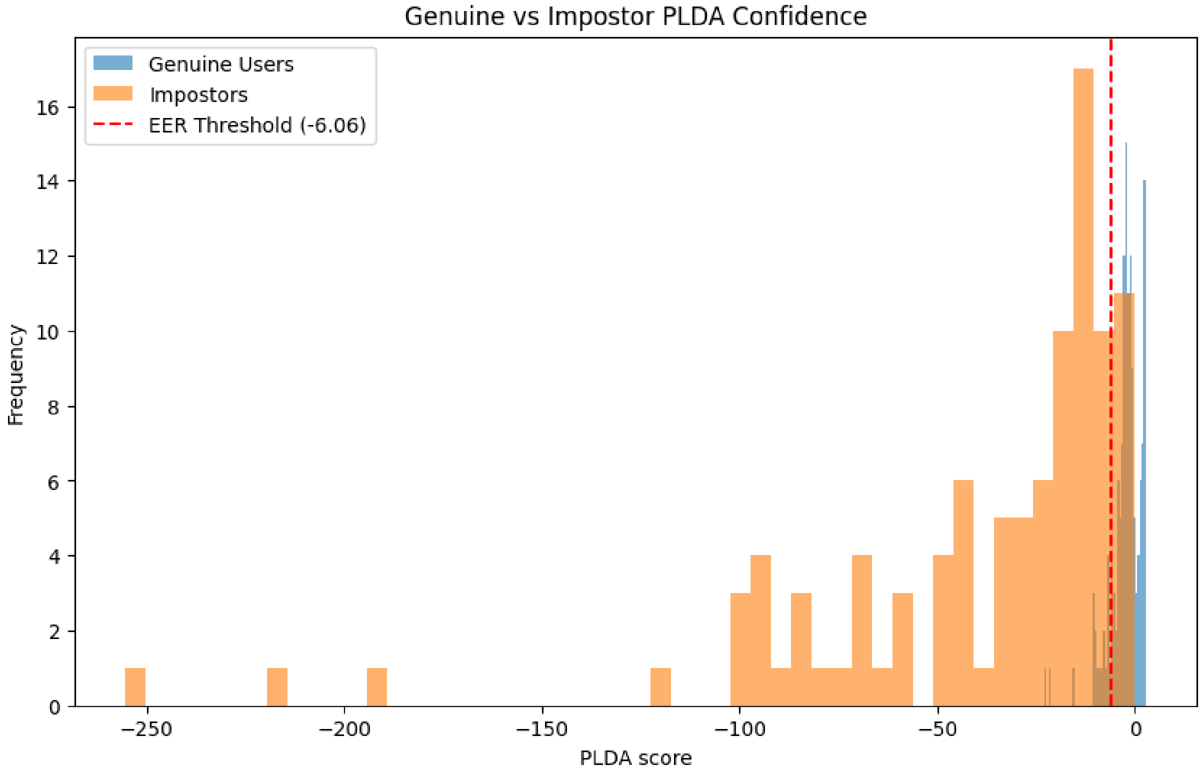
Feature Extraction		*Speaker Identification*	*Speaker Verification*		*Metrics*
Features used	**Frame Analysis**	**Model Used**	**Embedding**	**Backend**	**EER**
25 MFCCs	25ms frame length 10ms frame hop	DNN	Bottle-neck layer	GMM-UBM	0.03
25 MFCCs	25ms frame length 10ms frame hop	DNN	Last Hidden layer	GMM-UBM	0.13
25 MFCCs	25ms frame length 10ms frame hop	RNN	Mean Hidden Layers	GMM-UBM	0.12
25 MFCCs	25ms frame length 10ms frame hop	CNN+LSTM	Mean Hidden Layers	GMM-UBM	0.15
20 MFCCs	25ms frame length 10ms frame hop	I-DNN	I-embedding	GMM-UBM	0.16
24 MFCCs, 24 Delta-MFCCs VAD 30% energy	25ms frame length 10ms frame hop	TDNN	X-embedding	GMM-UBM	0.48

Table 5.3: Speaker Verification

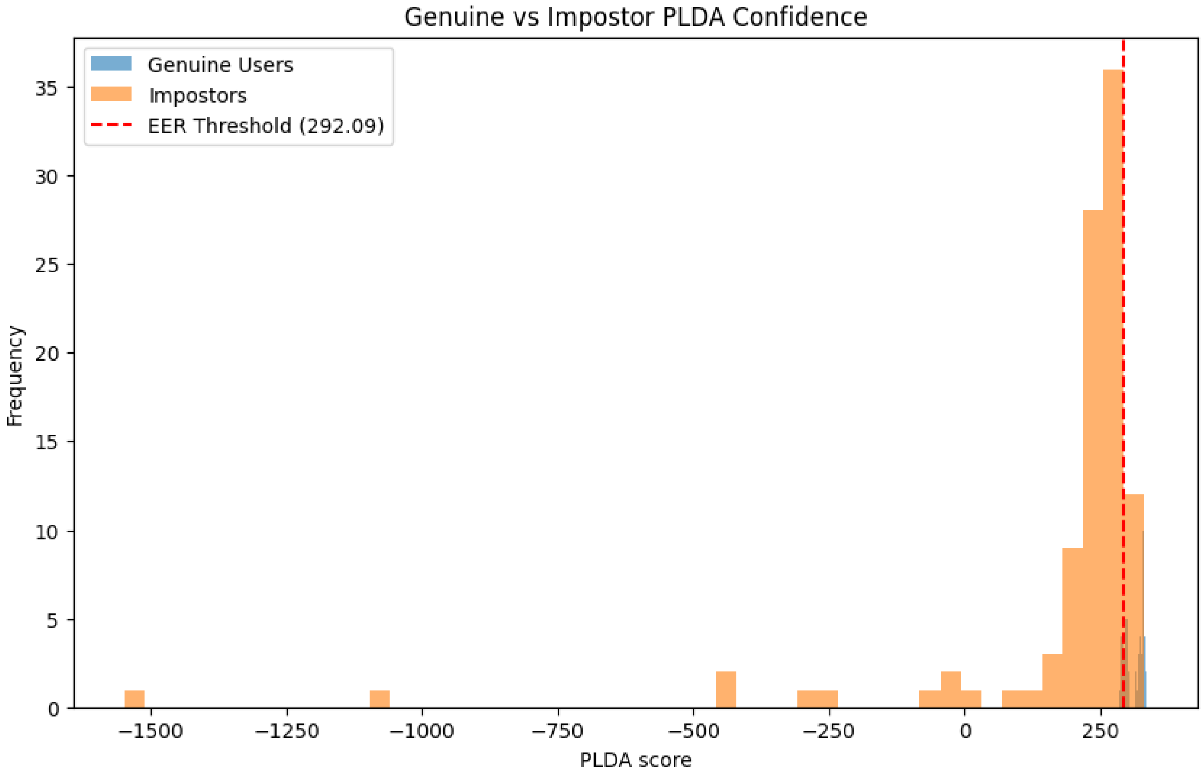
Architettura: DNN-1



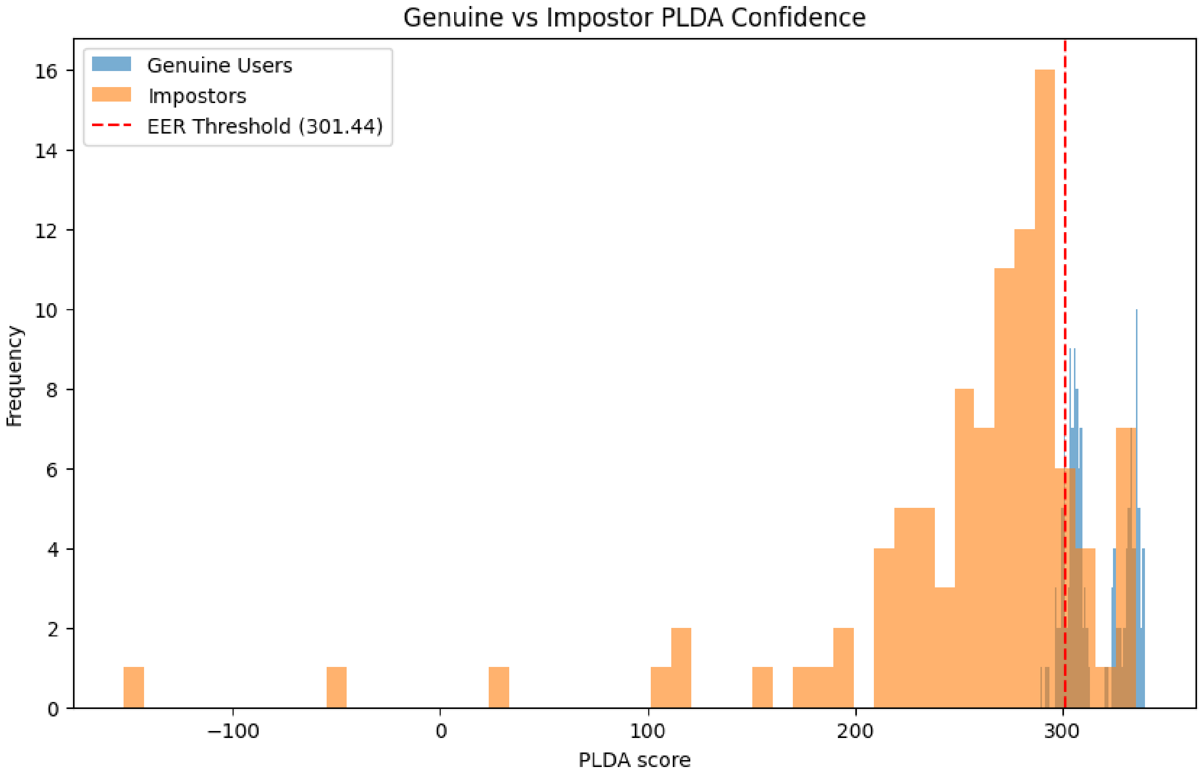
Architettura: DNN-2



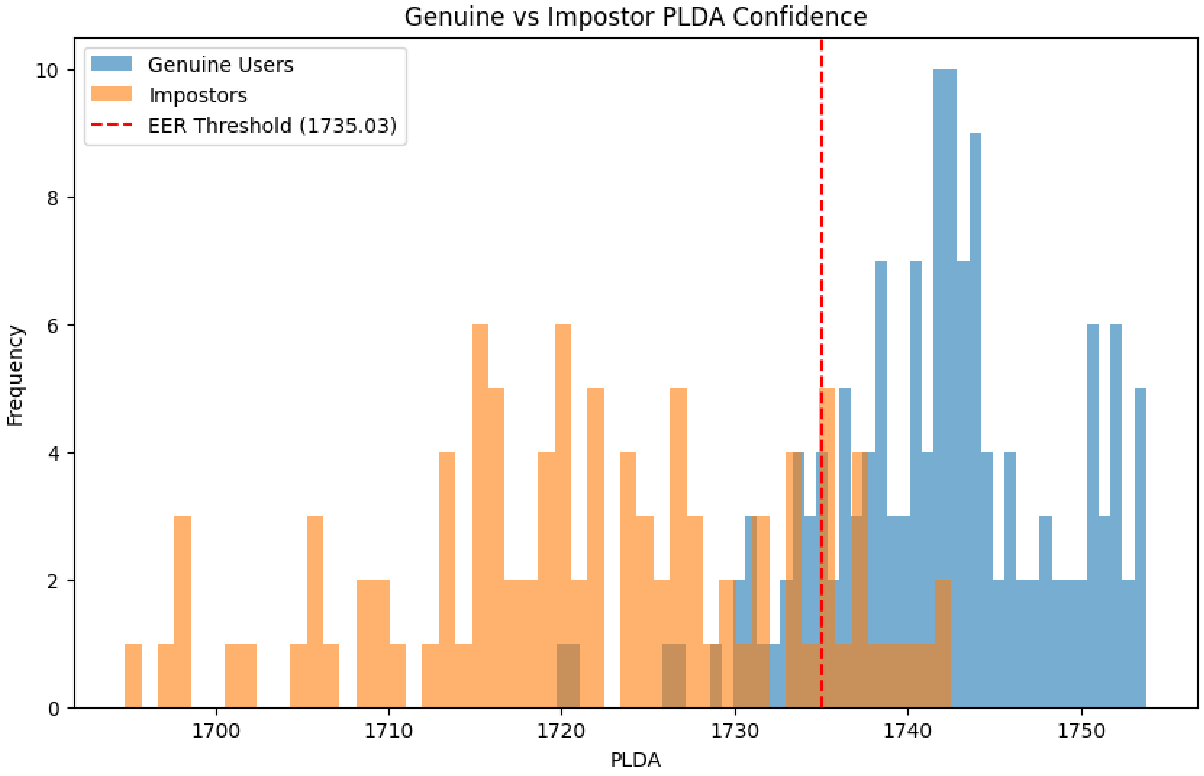
Architettura: RNN



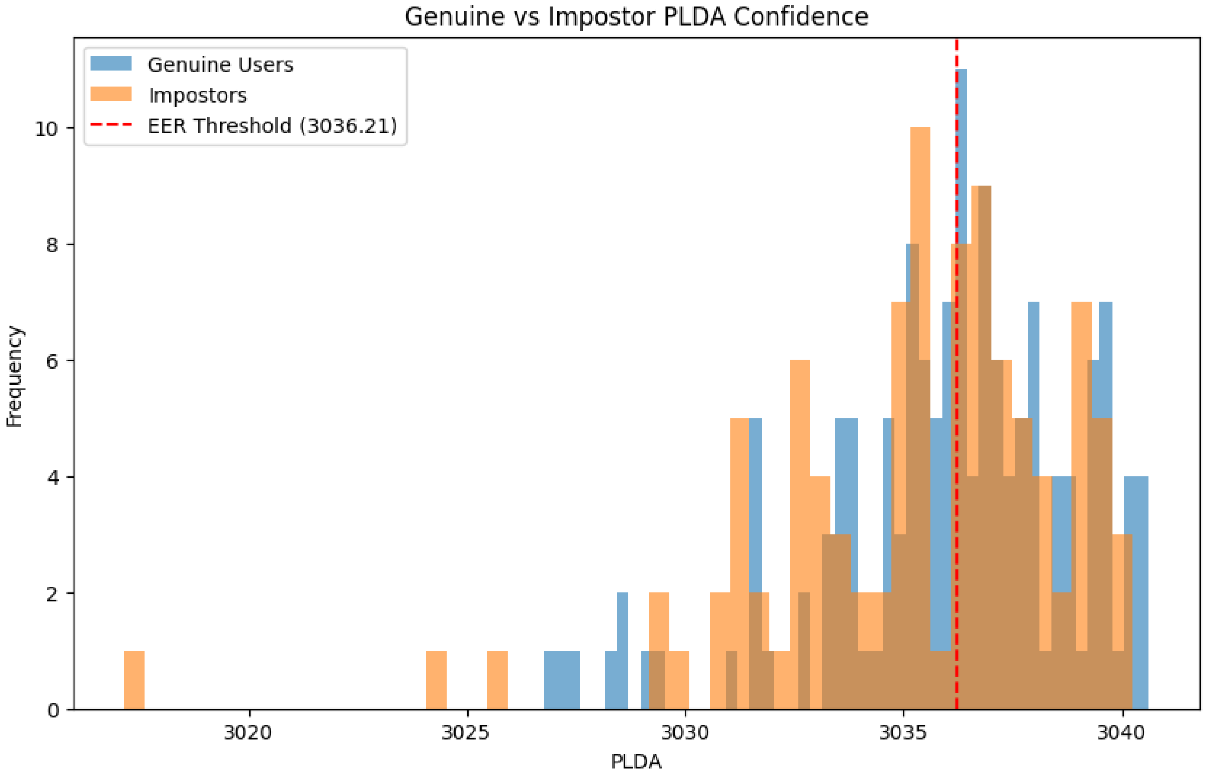
Architettura: CNN+LSTM



Architettura: I-DNN



Architettura: X-DNN


Genuine vs Impostor PLDA Confidence
Genuine Users
Impostors
EER Threshold (3036.21)
Frequency
PLDA