A Corrective Learning Approach For Text-Independent Speaker Verification

ICASSP 2018, Calgary, Canada

Yandong Wen, Tianyan Zhou, Rita Singh*, and Bhiksha Raj

Carnegie Mellon University

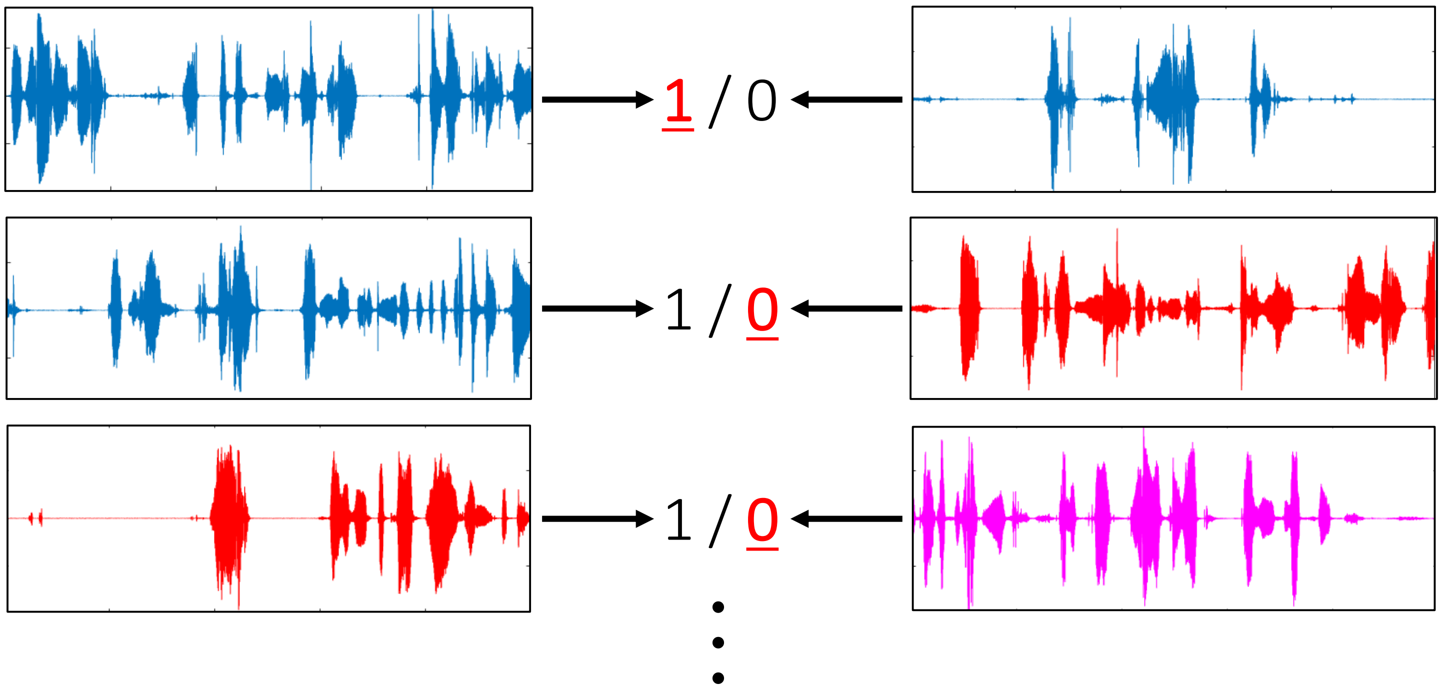
Introduction

Task

Determining if the speaker in a "test" recording is the same as that in a prior "enrollment" recording

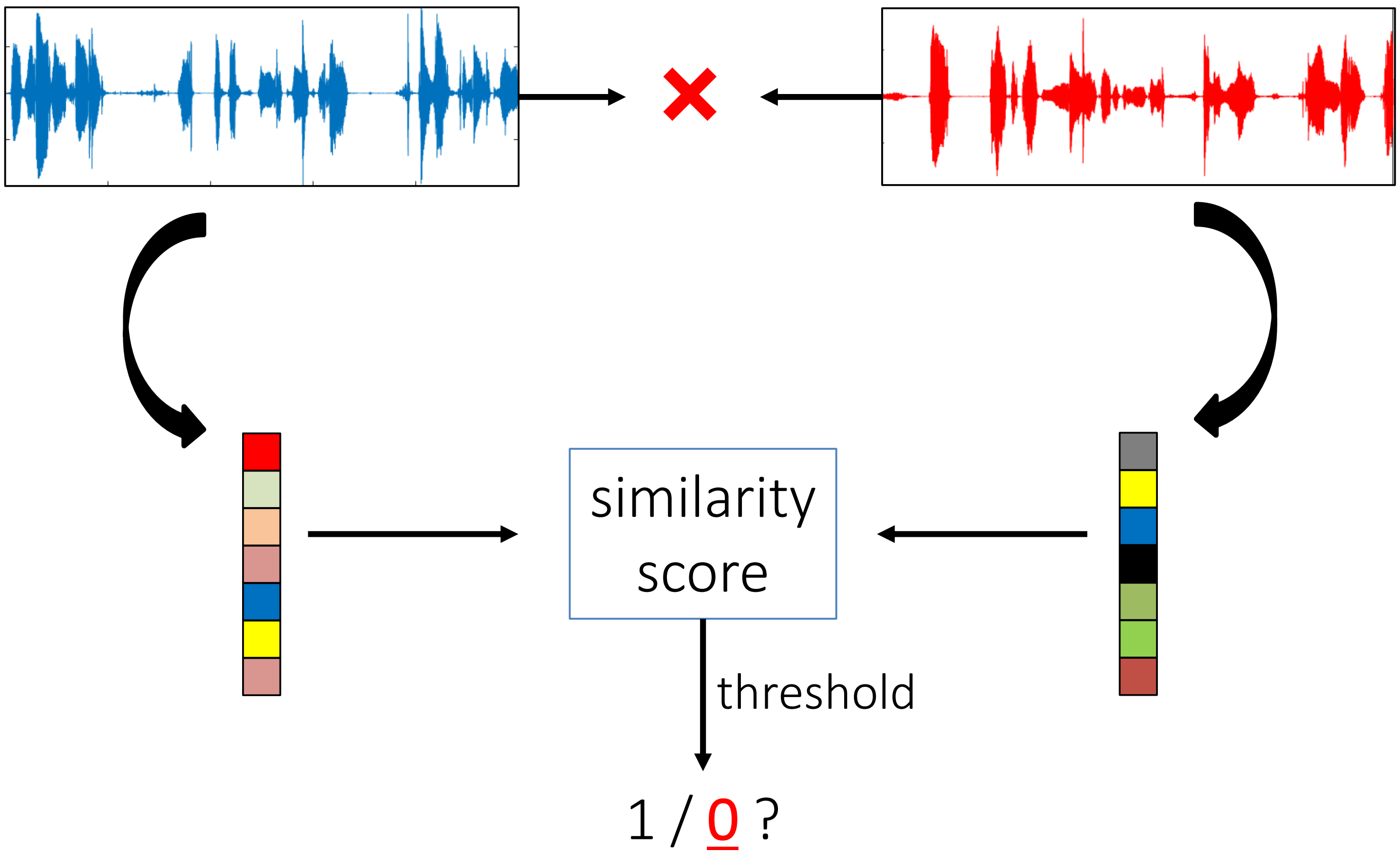
Recordings in Enrollment

Recordings in Test



1: positive pair   0: negative pair

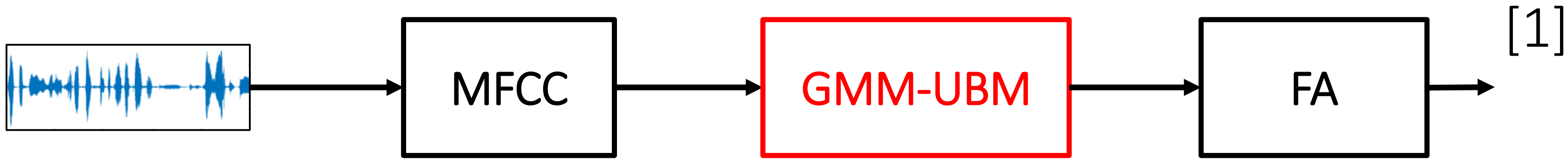
Representations



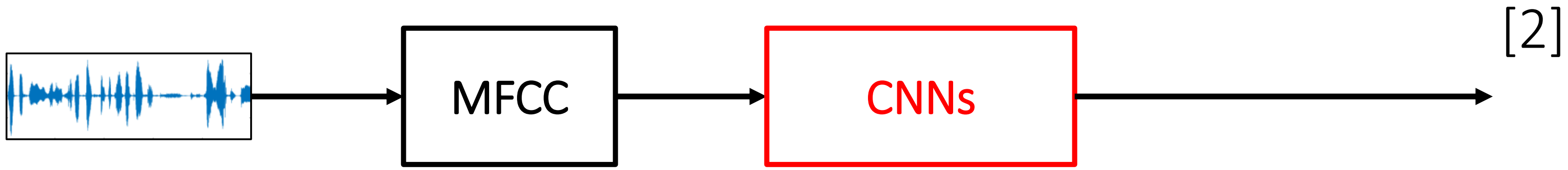
Prior work

Variable-length recordings are represented as fixed-length vectors

i-vector system



CNN system



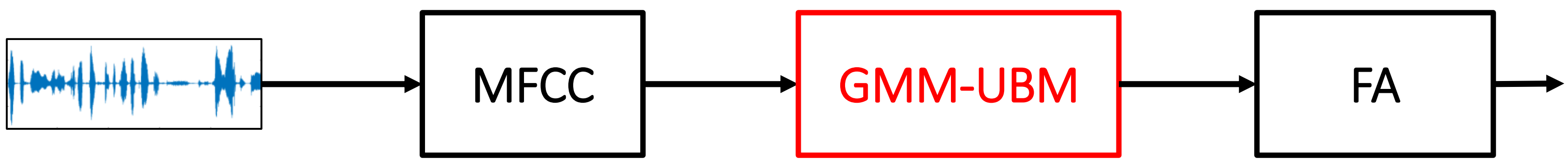
[1] Dehak, N., et al. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), 788-798.

[2] Snyder, D., et al. Deep neural network-based speaker embeddings for end-to-end speaker verification. In Spoken Language Technology Workshop (SLT), 2016 IEEE (pp. 165-170). IEEE.

Prior work

i-vector system

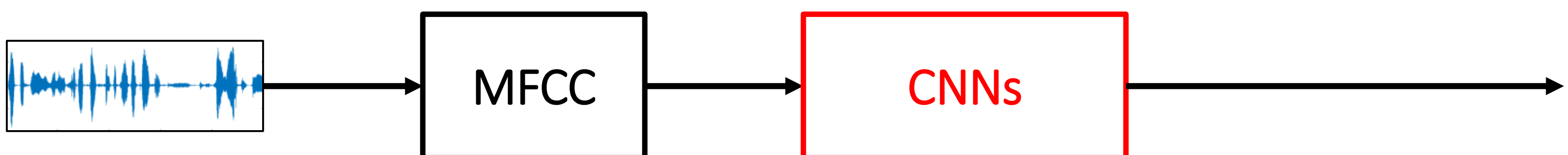
Supervector: concatenation of the **<u>means</u>**



[1]

CNN system

Temporal pooling: **<u>average</u>** across time domain



[2]

averaging

Proposed method

Formulation

input: $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$ is a collection of speech segments from a recording with class Y

objective: $\hat{Y} = \arg\max_{Y} P(Y|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$

$$\prod_i P(Y|\boldsymbol{x}_i) \neq P(Y|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$$

Taking average is **NOT** perfectly reasonable

Even if $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$ are class-conditionally independent

Alternative perspective

objective: $\hat{Y} = \arg\max_{Y} P(Y|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$

$$P(Y|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t) = \frac{P(Y|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{t-1})P(\boldsymbol{x}_t|Y)}{P(\boldsymbol{x}_t|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{t-1})}$$

ignorable

Log on both sides

$$L_{t-1}(Y) = \log P(Y|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{t-1})$$
$$\Delta L(Y, \boldsymbol{x}_t) = \log P(\boldsymbol{x}_t|Y)$$

P: probability
L: log likelihood

$$\hat{Y}_t = \arg\max_{Y} L_{t-1}(Y) + \Delta L(Y, \boldsymbol{x}_t)$$

previous prediction correction

Electrical & Computer ENGINEERING

Incremental Bayesian classification

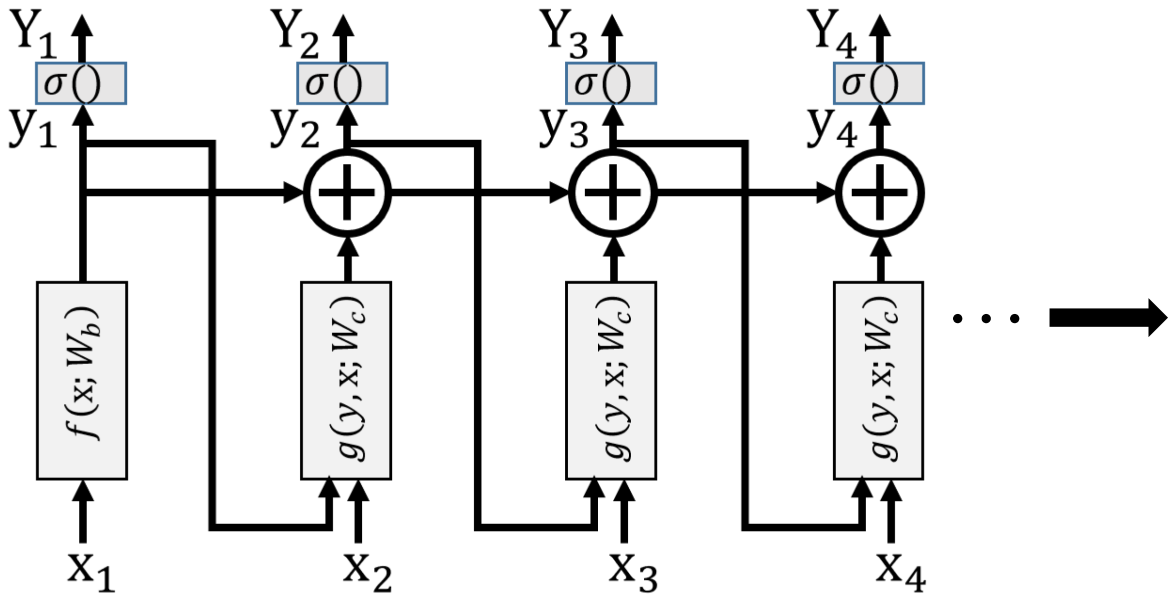
objective: $\hat{Y} = \arg\max_{Y} P(Y|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$

$$\hat{Y}_t = \arg\max_{Y} L_{t-1}(Y) + \Delta L(Y, \boldsymbol{x}_t)$$

previous prediction correction

- Speech segments x_1, x_2, ..., x_n are assumed to be conditionally independent and orderless.
- Use new speech segments x_1, x_2, ..., x_n to build upon the predictions that already made.
- This recurrent formalism is called deep corrective learning networks (CLNets)

Deep corrective learning nets



$$\boldsymbol{y}_1 = f(\boldsymbol{x}_1; W_f)$$

$$\Delta \boldsymbol{y}_t = g(\boldsymbol{y}_{t-1}, \boldsymbol{x}_t; W_g)$$

$$\boldsymbol{y}_t = \boldsymbol{y}_{t-1} + \Delta \boldsymbol{y}_t$$

$$Loss(Y, Y_N) = \sum_{t=1}^{N} w_t Loss(Y, Y_t)$$

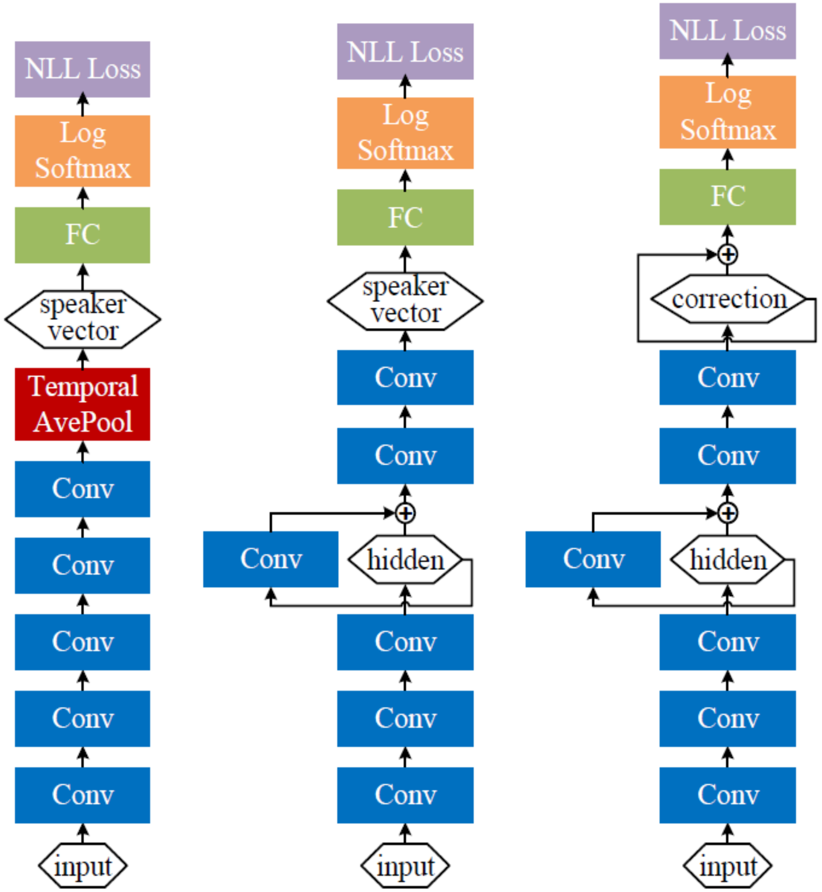
Experiments

Datasets

- Training data: SRE04 – 08
  - 36,500 recordings, 3801 speakers, 5 mins
- Testing data: SRE10
  - 11,959 recordings for enrollment, 5 mins
  - 767 recordings for testing, 5 mins
  - Trial file: 416,119 pairs
  - 7,169 positive pairs & 408,950 negative pairs

Network architecture

- 5 convolutional layers.
- # of filters: 4, 16, 64, 256, 64
- Filter size: 3x3
- Stride: 2
- Padding: 0
- Feature dimension: 64

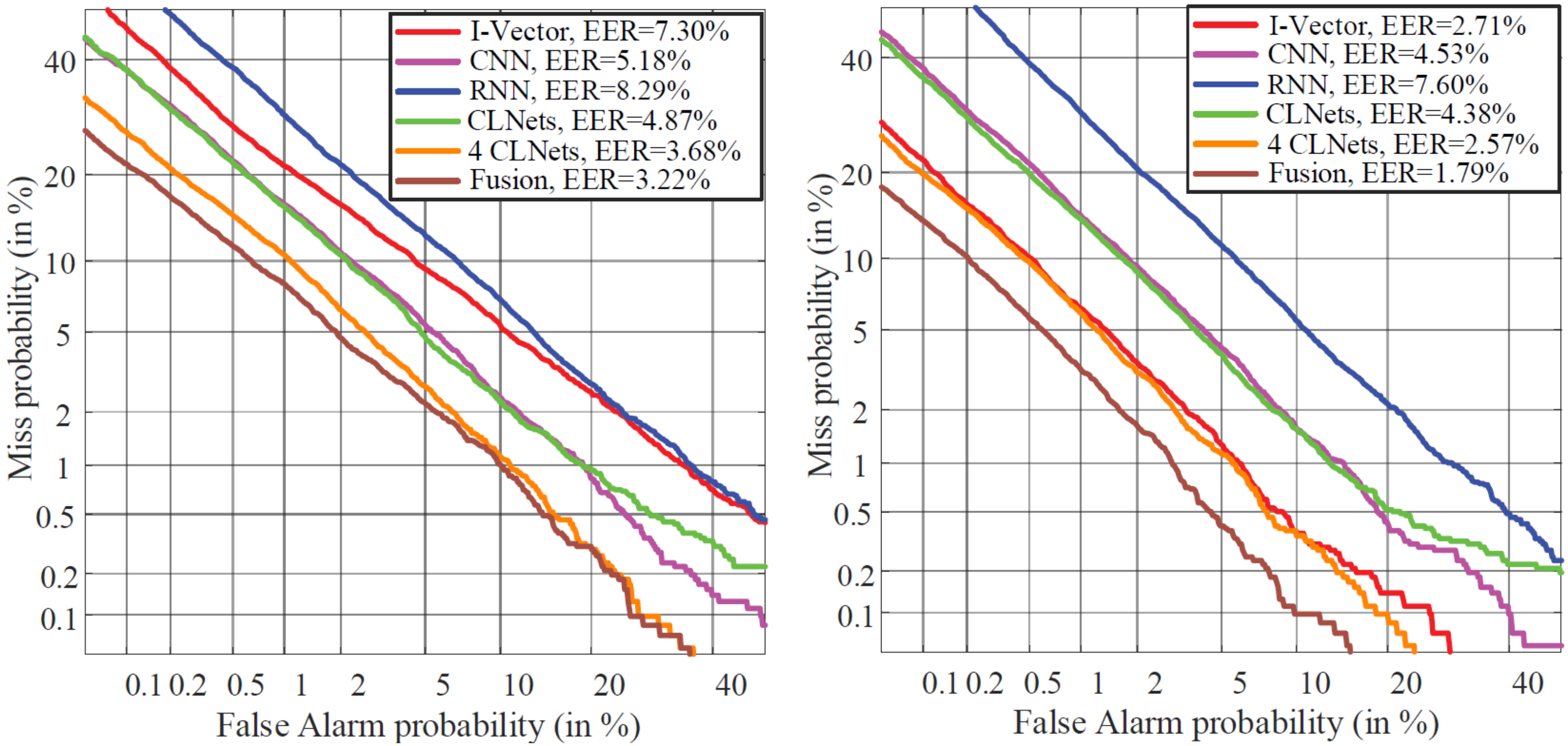


(a) CNNs (b) RNNs (c) CLNets

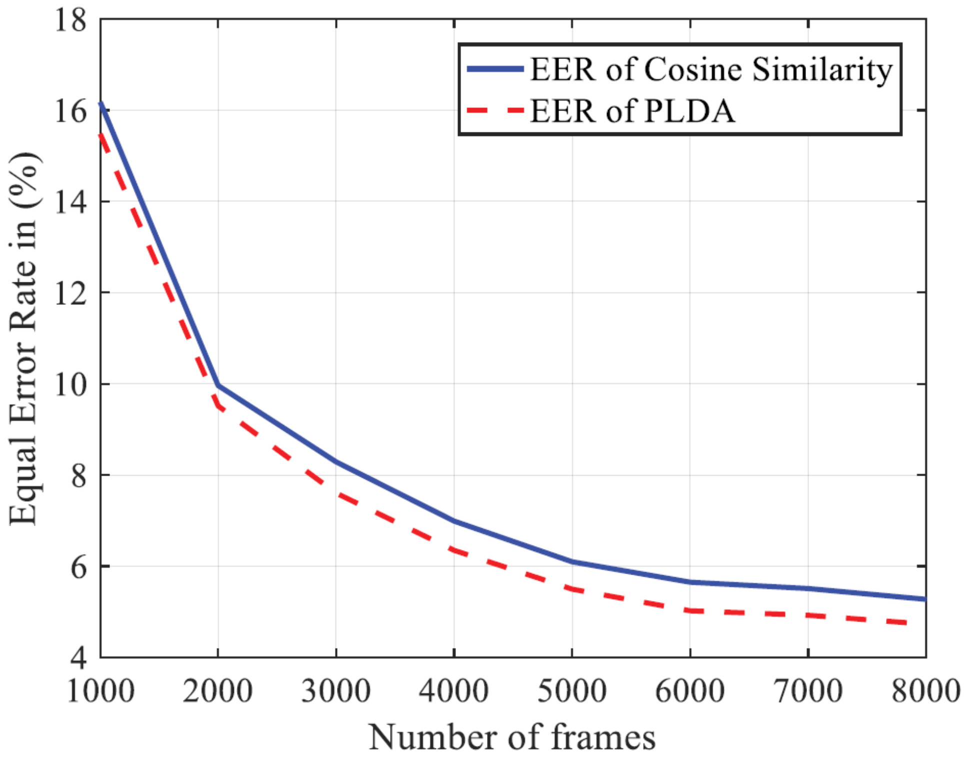
Evaluation

- The extended core condition 5 on SRE10 (7,169 positive pairs & 408,950 negative pairs)
  - Entire recordings
  - Enrollment and testing recordings are truncated from 10 to 80 seconds with a granularity of 10 seconds
- Score computation:
  - Cosine Similarity and PLDA
- Performance measurement
  - Detection error tradeoff (DET) curves and equal error rates (EER)

Performance



Performance



Thank You!