Machine Learning for Speaker Recognition

Sandro Cumani

Politecnico di Torino

Outline

- Speaker Identification and Verification:
  - The Speaker Verification Chain
  - Voiceprint Classification - PSVM
  - Voiceprint Extraction
  - Generative Classification - PLDA extensions
- Log-Likelihood Ratios and Score Calibration
- Speaker diarization and Clustering
- Beyond Voice Biometrics: Face ID, multi- and cross-modal identification

Speaker Recognition

Speaker Recognition:

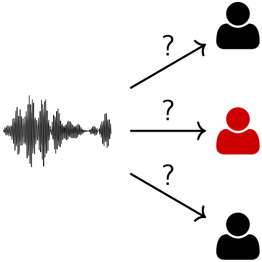
- Authentication
- Surveillance
- Forensics

Strong territorial presence in Torino:

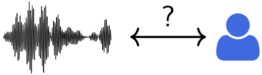
- Telecom $\rightarrow$ Loquendo $\rightarrow$ Nuance Communications
- Research contracts, joint participation to NIST Evaluations

Speaker Recognition

Speaker identification:
who is speaking?

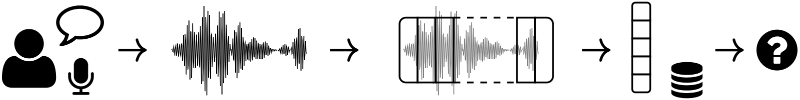


Speaker verification:
is A speaking?



The Speaker Verification Chain

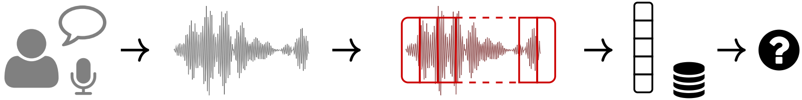
Classical approach to Speaker Verification:



- Acoustic feature extraction
- Voiceprint extraction
- Voiceprint classification

The Speaker Verification Chain

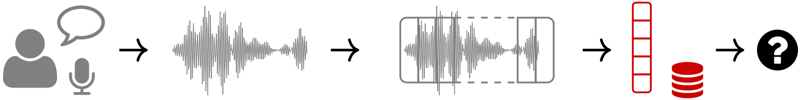
Classical approach to Speaker Verification:



- Acoustic feature extraction
- Voiceprint extraction
- Voiceprint classification

The Speaker Verification Chain

Classical approach to Speaker Verification:



- Acoustic feature extraction
- Voiceprint extraction
    - Statistical: i-vectors
    - Neural: x-vectors
- Voiceprint classification

The Speaker Verification Chain

From acoustic frames to voiceprints: i-vectors

- State-of-the-art for the last decade ($\approx$ 2010 to $\approx$ 2018)
- Still the best solution for some tasks

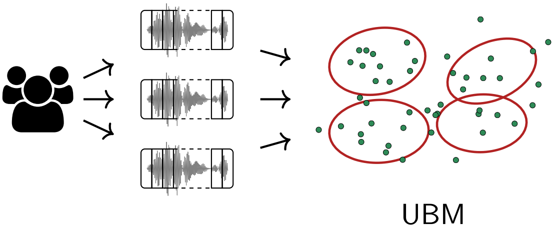
Low-dimensional, MAP-adapted, GMM supervector

The Speaker Verification Chain

From acoustic frames to voiceprints: i-vectors

- State-of-the-art for the last decade ($\approx$ 2010 to $\approx$ 2018)
- Still the best solution for some tasks

Low-dimensional, MAP-adapted, GMM supervector



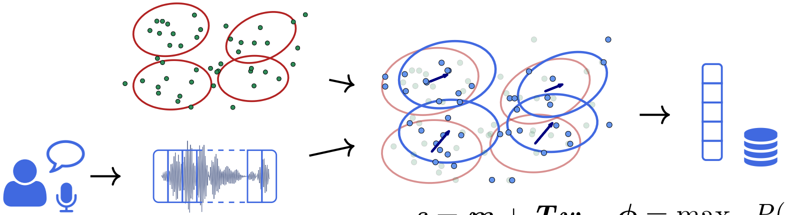
$$P(\boldsymbol{x}_t) = \sum_c w_c \mathcal{N}(\boldsymbol{x}_t | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

The Speaker Verification Chain

From acoustic frames to voiceprints: i-vectors

- State-of-the-art for the last decade ($\approx$ 2010 to $\approx$ 2018)
- Still the best solution for some tasks

Low-dimensional, MAP-adapted, GMM supervector



$\mathcal{X}$

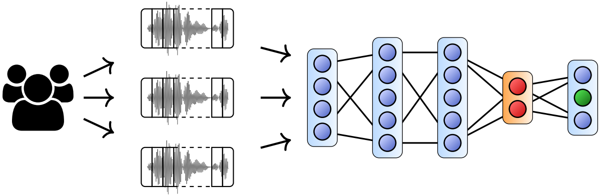
$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \qquad \phi = \max_w P(\boldsymbol{w}|\mathcal{X})$$

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

The Speaker Verification Chain

From acoustic frames to voiceprints: neural networks

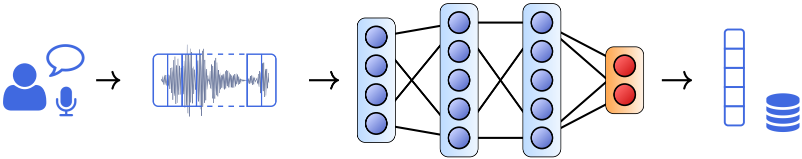
- Current state-of-the-art utterance representation
- Non-linear embedding transformations trained for multiclass classification (cross-entropy)



The Speaker Verification Chain

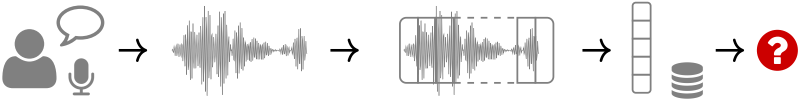
From acoustic frames to voiceprints: neural networks

- Current state-of-the-art utterance representation
- Non-linear embedding transformations trained for multiclass classification (cross-entropy)



The Speaker Verification Chain

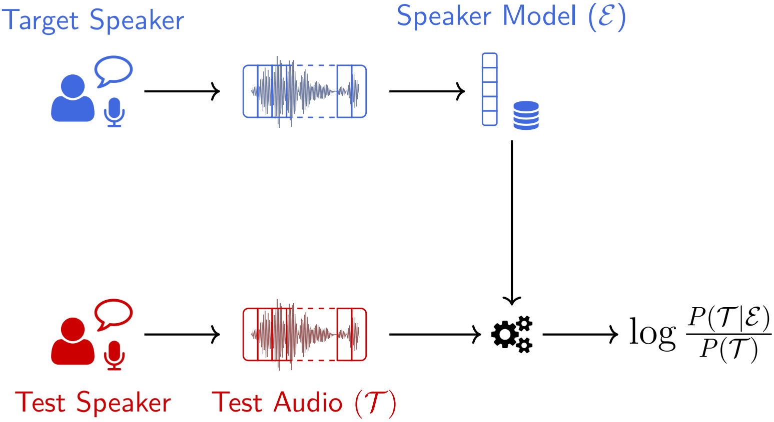
Classical approach to Speaker Verification:



- Acoustic feature extraction
- Voiceprint extraction
- Voiceprint classification

Voiceprint Classification

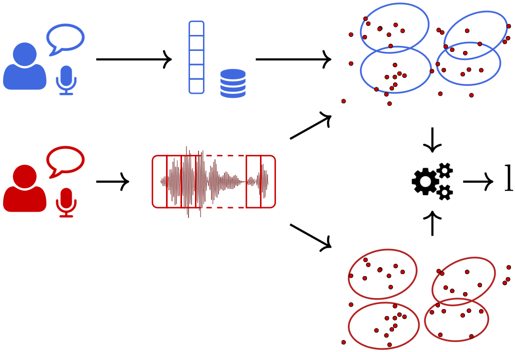
Voiceprint vs Audio: is segment $\mathcal{T}$ from speaker $\mathcal{E}$?



Voiceprint Classification

Voiceprint vs Audio: Generative approach (pre-2010)

- Channel-compensated GMM adapted supervector per speaker
- Compute likelihood for test frames

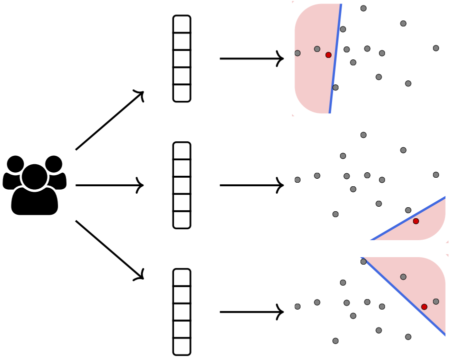


$$\to \log \frac{P(\mathcal{T}|\mathcal{E})}{P(\mathcal{T})}$$

Voiceprint Classification

Voiceprint vs Audio: Discriminative (SVM) approach:

- Train 1-vs-all hyperplane in i-vector space



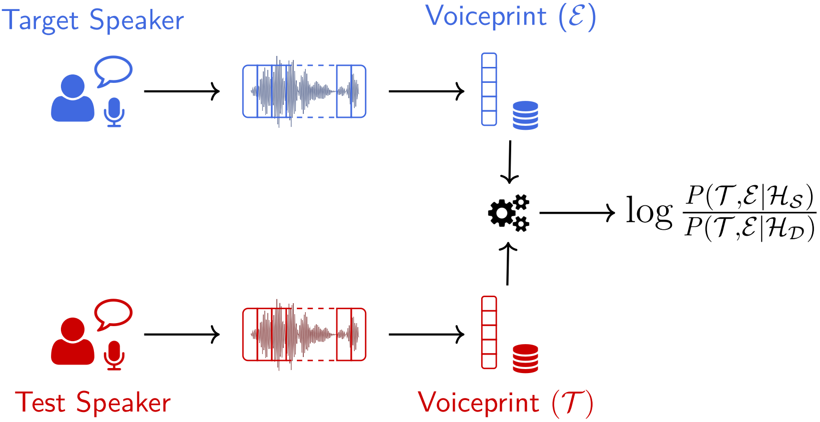
Voiceprint Classification

Voiceprint vs Audio: Discriminative (SVM) approach:

- Train 1-vs-all hyperplane in i-vector space
- Very unbalanced classes
- Few (single) points for target
- Each model has different score dynamics

Voiceprint Classification

Voiceprint vs voiceprint: are $\mathcal{T}$ and $\mathcal{E}$ from the same speaker?

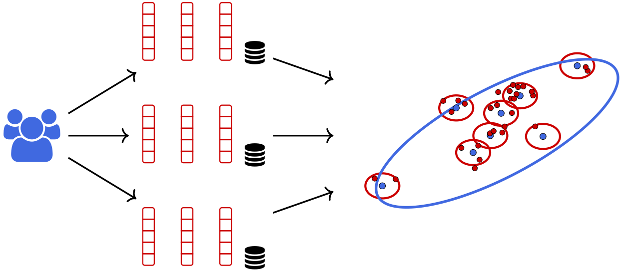


Voiceprint Classification

Voiceprint vs Voiceprint: Generative approach

- Model between-class and within-class variability
- State-of-the-art: PLDA (and variations):

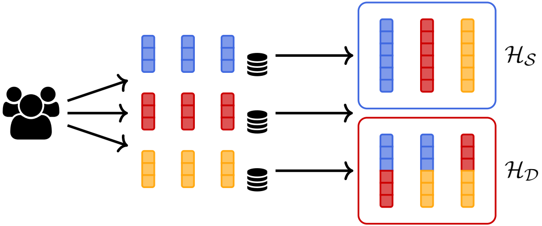
$$\boldsymbol{\phi}_{s,i} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{y}_s + \boldsymbol{\varepsilon}_{s,i} \ , \quad \boldsymbol{y}_s \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \ , \quad \boldsymbol{\varepsilon}_{s,i} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Lambda}^{-1}\right)$$



Voiceprint Classification

Voiceprint vs Voiceprint: Pairwise SVM[1][2]

- Classify pairs of i-vectors (trials)
- Binary problem: Same- vs Different-speaker trial



[1] S. Cumani et al. "Pairwise Discriminative Speaker Verification in the I-Vector Space". In: IEEE Transactions on Audio, Speech, and Language Processing 21.6 (2013), pp. 1217–1227.

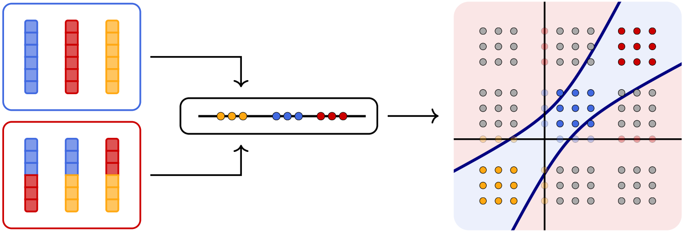
[2] S. Cumani and P. Laface. "Large scale training of Pairwise Support Vector Machines for speaker recognition". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 22.11 (2014), pp. 1590–1600.

Voiceprint Classification

Voiceprint vs Voiceprint: Pairwise SVM

- Single model trained on background speakers
- Quadratic Kernel (similar to PLDA Log-Likelihood Ratios)

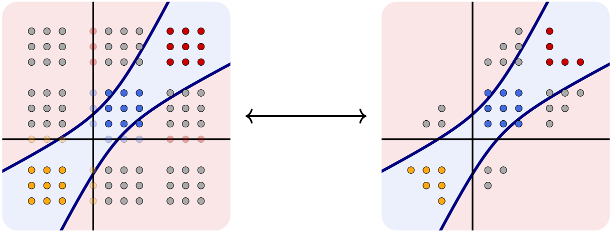
$$s(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{y} + \boldsymbol{y}'\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} + \boldsymbol{y}'\boldsymbol{B}\boldsymbol{y} + \boldsymbol{c}'\boldsymbol{x} + \boldsymbol{c}'\boldsymbol{y} + k$$



Voiceprint Classification

Voiceprint vs Voiceprint: Pairwise SVM

- N utterances $\rightarrow$ N^2 pairs $\rightarrow$ Primal solver
- Naive approach: $O(N^2D^2)$ per iteration (expanded features)
- Exploit score correlations: $O(N^2D + ND^2)$ per iteration
- Support Vector Filtering: $O(ND^2)$ per iteration

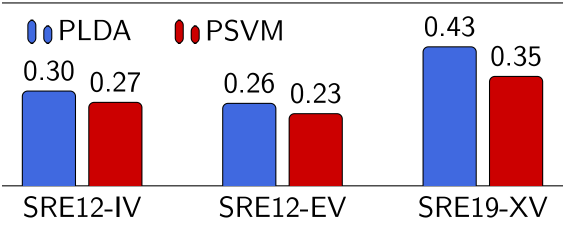


Voiceprint Classification

Voiceprint vs Voiceprint: Pairwise SVM

- More accurate, scoring costs as PLDA, both can be combined
- Effective with different front-ends (i-/e-[3]/x-vector)

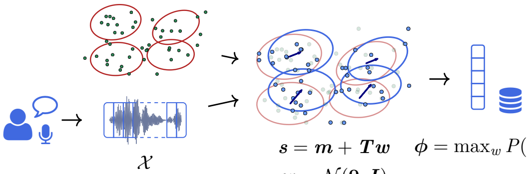
Decision Costs (lower is better)



[3] S. Cumani and P. Laface. "Speaker recognition using e–vectors". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.4 (2018), pp. 736–748.

Voiceprint extraction: i-vectors

Low-dimensional, MAP-adapted, GMM supervector



$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \qquad \boldsymbol{\phi} = \max_{w} P(\boldsymbol{w}|\mathcal{X})$$

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

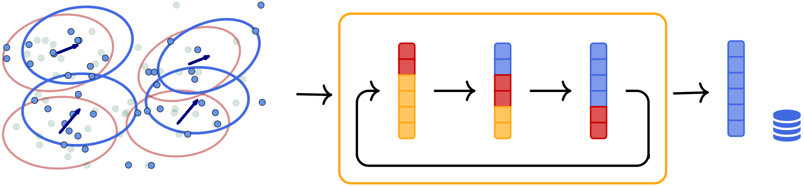
- Computationally expensive:
  - Fast / large memory or very slow / average memory
  - Not suited for embedded devices
- Simultaneous diagonalization[4]:
  - Very fast, average memory, significant accuracy degradation

[4] O. Glembek et al. "Simplification and optimization of i-vector extraction". In: Proceedings of ICASSP 2011.

Voiceprint extraction: i-vectors

Variational Bayes approximation of posterior distribution[5]

- Iterative (≈ fast as standard), average memory, same accuracy

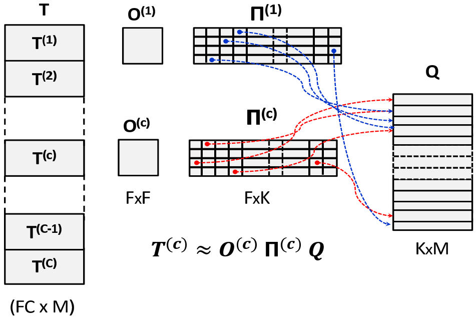


[5] S. Cumani and P. Laface. "Memory and computation trade-offs for efficient i-vector extraction". In: IEEE Transactions on Audio, Speech, and Language Processing 21.5 (2013), pp. 934–944.

Voiceprint extraction: i-vectors

Subspace Factorization (FSE)[6]

- Iterative (very fast), very low memory, limited accuracy loss



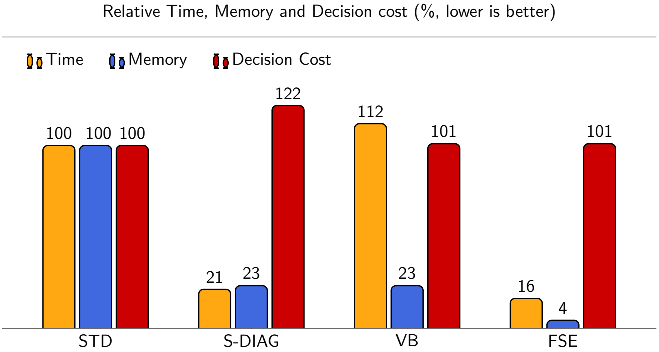
$$T^{(c)} \approx O^{(c)}\, \Pi^{(c)}\, Q$$

[6] S. Cumani and P. Laface. "Factorized Sub-Space Estimation for Fast and Memory Effective I-vector Extraction". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 22.1 (2013), pp. 248–259.

Voiceprint extraction: i-vectors

Subspace Factorization

- Best solution for low-resource devices (Patent[7])

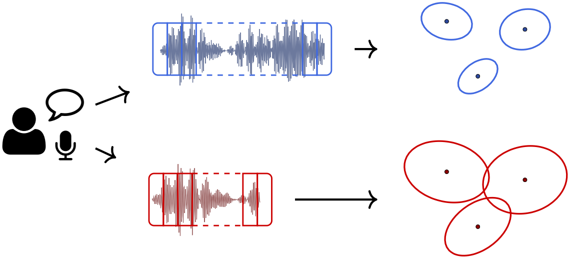


[7] Sandro Cumani and Pietro Laface. "Method and apparatus for efficient i–vector extraction". Patent. US 9406298 B2. 2016.

PLDA for short utterances

I-vectors and short utterances:

- Shorter utterance $\rightarrow$ less accurate i-vector estimation
- Uncertainty can be quantified (i-vector posterior)

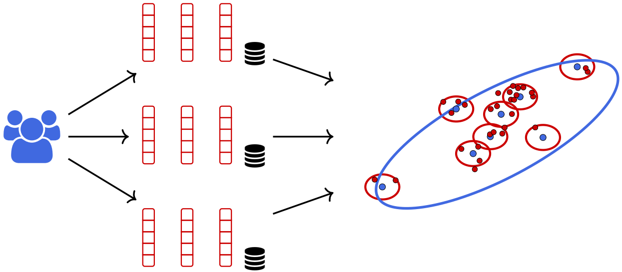


PLDA for short utterances

Natural framework for modeling uncertainty: PLDA $\rightarrow$ FPD-PLDA

- Complementary to PSVM
- Probabilistic generative model for i-vectors

$$\boldsymbol{\phi}_{s,i} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{y}_s + \boldsymbol{\varepsilon}_{s,i} \ , \qquad \boldsymbol{y}_s \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \ , \qquad \boldsymbol{\varepsilon}_{s,i} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Lambda}^{-1}\right)$$



PLDA for short utterances

Natural framework for modeling uncertainty: PLDA → FPD-PLDA

- Integrate over uncertainty in PLDA model[8] (FPD-PLDA)
- Higher accuracy, much slower
- Approximate solutions for fast scoring[9] (Patent[10])

---

[8] S. Cumani, O. Plchot, and P. Laface. "On the use of i-vector posterior distributions in Probabilistic Linear Discriminant Analysis". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 22.4 (2014), pp. 846–857.

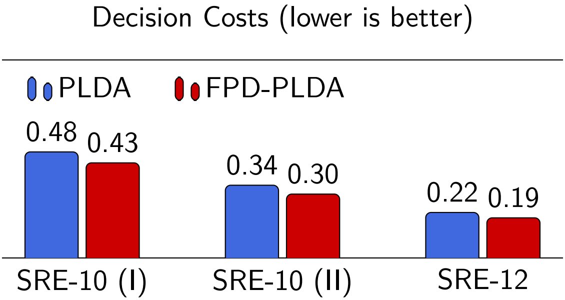
[9] S. Cumani. "Fast Scoring of Full Posterior PLDA Models". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 23.11 (2015), pp. 2036–2045.

[10] Sandro Cumani et al. "Fast speaker recognition scoring using i–vector posteriors and Probabilistic Linear Discriminant Analysis". Patent. US 9373330 B2. 2016.

PLDA for short utterances

Natural framework for modeling uncertainty: PLDA → FPD-PLDA

- Very good results for ABC[11] team in NIST SRE 2012
- Very effective for short and variable duration utterances



[11] Agnitio, Brno University of Techology, CRIM

Non-Linear PLDA

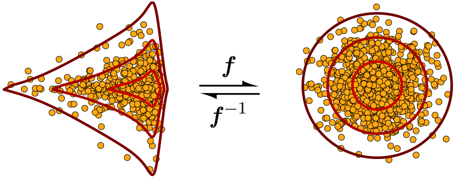
Beyond Gaussian assumptions: Non-Linear PLDA

- PLDA: Gaussian-Linear model
- Remove Gaussian Assumptions
- Limit the impact on scoring time
  - Closed form, simple posteriors

Non-Linear PLDA

Non-Linear PLDA: Combine PLDA and density transformations[12][13]

- Transform input space to better match Gaussian assumptions
- Invertible function f
- Corresponds to non-Gaussian density in original space



[12] S. Cumani and P. Laface. "Non–linear i–vector transformations for PLDA based speaker recognition". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.4 (2017), pp. 908–919.

[13] S. Cumani and P. Laface. "Joint estimation of PLDA and non–linear transformations of speaker vectors". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.10 (2017), pp. 1890–1900.

Non-Linear PLDA

Non-Linear PLDA: Combine PLDA and density transformations

- Apply non-linear, invertible transformation to (i/e/x)-vectors

$$\boldsymbol{\phi}_{s,i} = \boldsymbol{f}^{-1}\left(\boldsymbol{m} + \boldsymbol{U}\boldsymbol{y}_s + \boldsymbol{\varepsilon}_{s,i}\ , \boldsymbol{\vartheta}\right)$$

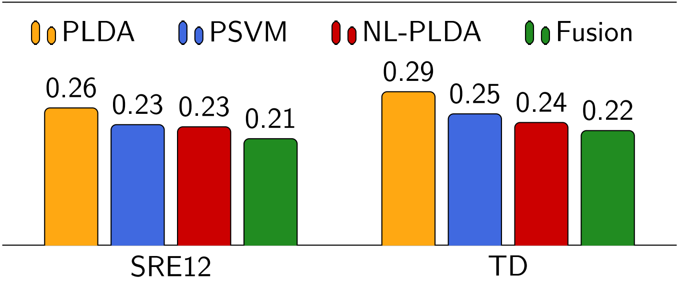
- Keep PLDA priors $\boldsymbol{y}_s \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $\quad \boldsymbol{\varepsilon}_{s,i} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{-1}\right)$
- Non-Gaussian likelihood, Conjugate Gaussian prior
  - Closed form, Gaussian posterior
  - Scoring complexity close to PLDA
- $\boldsymbol{f}$: composition of affine and $\sinh\text{-}\sinh^{-1}$ functions
- Training: EM + L-BFGS + modified back-prop

Non-Linear PLDA

Non-Linear PLDA: Combine PLDA and density transformations

- Improvement over PLDA, esp. for Text-Dependent (TD)
- Further gains from PSVM and NL-PLDA fusion

Decision Costs (lower is better)



Score calibration

Voiceprint vs Voiceprint: from audio to score

- Ideally, scores represent log-likelihood ratios (LLR)
- LLRs: optimal decision depends only on application-dependent priors and error costs
- In practice, scores $\neq$ LLRs:
  - Non-probabilistic classifiers (PSVM)
  - Incorrect model assumptions
  - Train / Test mismatch

Score calibration: recover LLR interpretation

Score calibration

Non-parametric calibration

- Isotonic regression

Discriminative score calibration

- Optimize logarithmic proper scoring rule
- Prior-weighted Logistic Regression[14]

Generative score calibration: model score generation process

- LLR of LLR is the LLR[15]
- Constrained vs. uncostrained densities

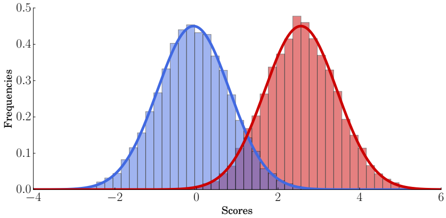
[14] N. Brümmer and G. R. Doddington. "Likelihood-ratio calibration using prior-weighted proper scoring rules". In: Proceedings of Interspeech. 2013, pp. 1976–1979.

[15] D. van Leeuwen and N. Brümmer. "The distribution of calibrated likelihood-ratios in speaker recognition". In: Proceedings of Interspeech. 2013, pp. 1619–1623.

Generative score calibration

Constrained ML calibration models

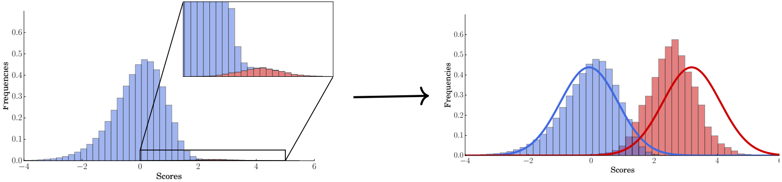
- Similar results to discriminative approaches
- Allow for semi- and unsupervised training
- Rely on assumptions on score distribution
  - E.g. CMLG → Tied-variance Gaussians



Generative score calibration

Unsupervised training:

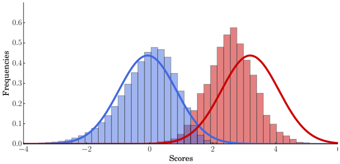
- Few targets, many non-targets
- Difficult to locate target scores
- Requires accurate modelling of distribution tails



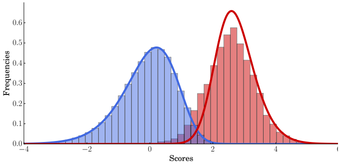
Generative score calibration

Unsupervised training:

- Theoretical (isotropic) PLDA LLRs: GH distributions[16]
- Contrained NIG[17][18] / VΓ / GH



CMLG



CMLGH

[16] Work in progress

[17] Sandro Cumani and Pietro Laface. "Tied Normal Variance–Mean Mixtures for Linear Score Calibration". In: Proceedings of ICASSP 2019. May 2019, pp. 6121–6125.

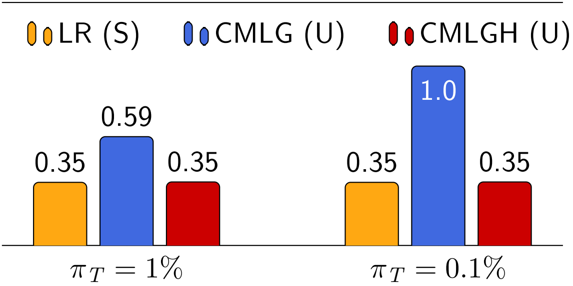
[18] Sandro Cumani. "Normal Variance-Mean Mixtures for Unsupervised Score Calibration". In: Proceedings of Interspeech 2019. Sept. 2019, pp. 401–405.

Generative score calibration

Unsupervised CMLGH:

- PLDA, NL-PLDA, PSVM, i-/e-/x-vectors
- Low target (speaker) proportion (1:1000 or lower)

Decision Costs (lower is better)

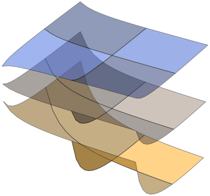


Generative score calibration

CMLGH score models: beyond calibration

- Model effects of distribution mismatch
- Compensate mismatch at score level
  - Single framework for score normalization and calibration
  - Incorporate duration effects for PSVM and x-vectors
  - Automatic determination of nuisance sources

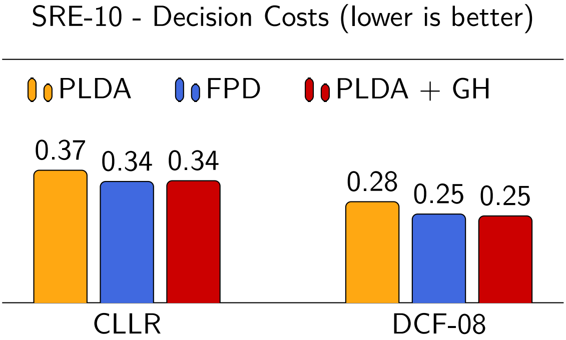
Duration-dependent score transformations:



Generative score calibration

CMLGH score models: beyond calibration

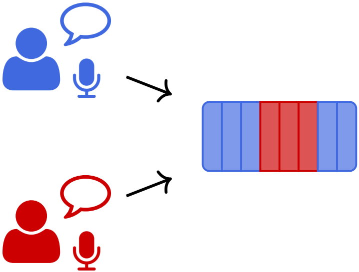
- Model effects of distribution mismatch
- Compensate mismatch at score level



Speaker diarization

Speaker segmentation and diarization:

- Single audio, multiple speakers
- Identify who is talking when



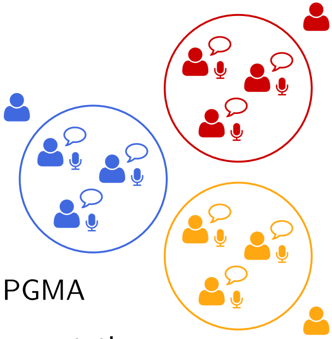
Speaker clustering

Cluster large number of speakers

- Fraudster detection
- Unsupervised adaptation

Large-scale UPGMA[19]

- Suited for SID similarities
- Constrained memory, fast, exact UPGMA
- Very fast, approximate silhouette computation

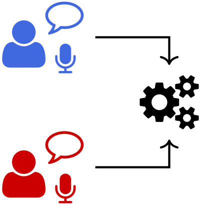


[19] Sandro Cumani and Pietro Laface. "Exact memory–constrained UPGMA for large scale speaker clustering". In: Pattern Recognition 95 (2019), pp. 235–246.

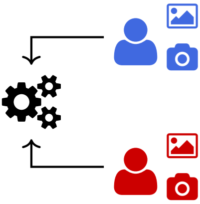
Face Identification

Speaker and Face Identification:

- Similar task (same / different identity)
- Different, but similar, frontends
  - i-vectors, x-vectors
  - DNN / CNN embeddings



$$\rightarrow \log \frac{P(\mathcal{T},\mathcal{E}|\mathcal{H}_{\mathcal{S}})}{P(\mathcal{T},\mathcal{E}|\mathcal{H}_{\mathcal{D}})} \leftarrow$$

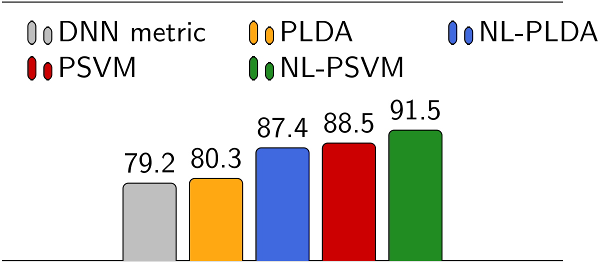


Face Identification

Speaker and Face Identification:

- NL-PLDA and PSVM effective[20]

SIFACE, Top-1 Accuracy (%, cross-age, higher is better)

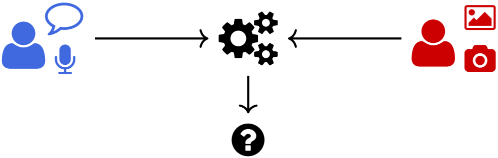


[20] Pablo Negri, Sandro Cumani, and Andrea Bottino. "PLDA-based Classification of Deep Features for Age-Invariant Face Recognition". Submitted to Computer Vision and Image Understanding.

Multi- and cross-modal Identification

Multi-modal and cross-modal identification

- Score-level, classifier, front-end fusion
- Cross-modal representation: joint face-voice embedding
- Multi and cross-modal backends: H-PLDA[21]



[21] Sandro Cumani and Pietro Laface. "Scoring Heterogeneous Speaker Vectors Using Nonlinear Transformations and Tied PLDA Models". In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2018).