Score Calibration
in speaker verification

Tianhao Wang

2023/03/17

Score Calibration

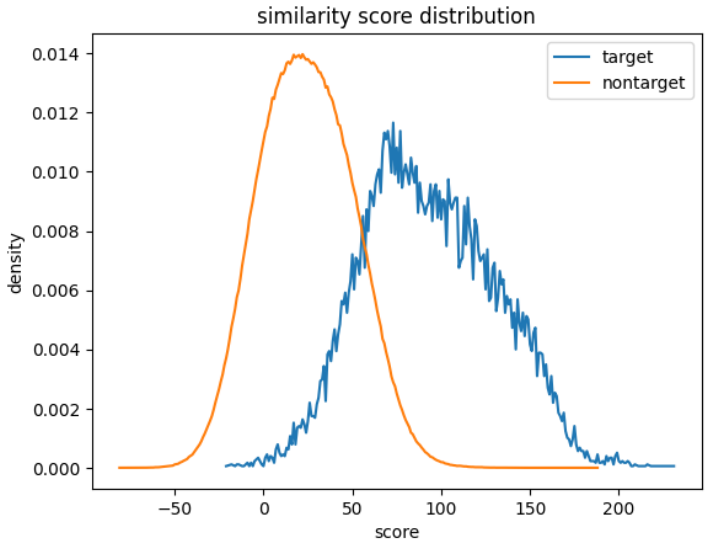
- Introduction
  - Mismatching between enrollment and test utterances leads to shifting of scores and unreliable accept/reject decisions.
  - Therefore, the purpose of score calibration is to suppress the impact of score drift on performance.

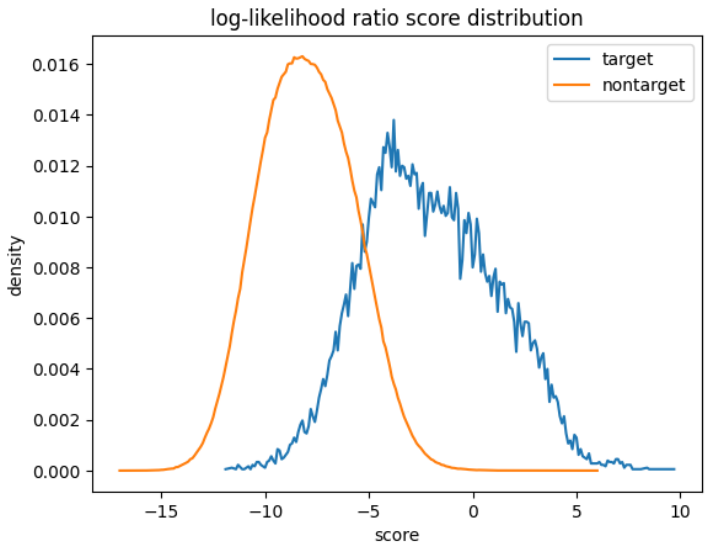
Linear Calibration:

- $y = l(s) = w_0 s + b$
- Converting **similarity score** to **log-likelihood ratio** $l(s) = log\frac{p_1(s)}{p_0(s)}$
- Fusion: $y = l(s) = \sum_{i=1}^{n} w_i s_i + b$

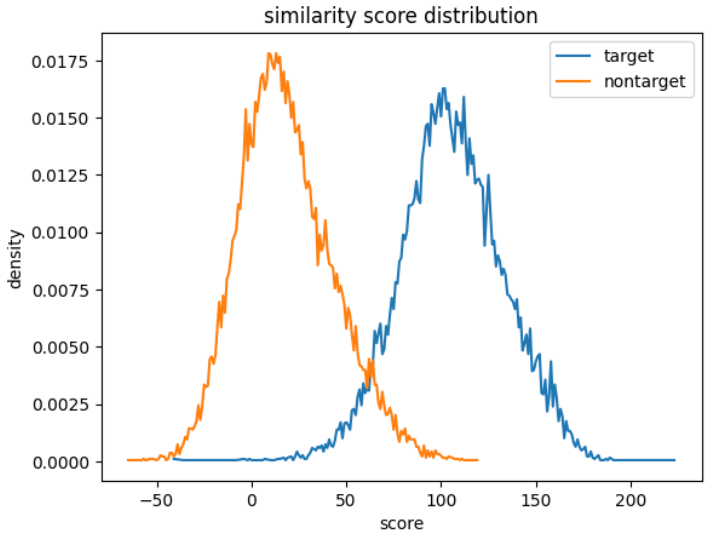
Score Distribution

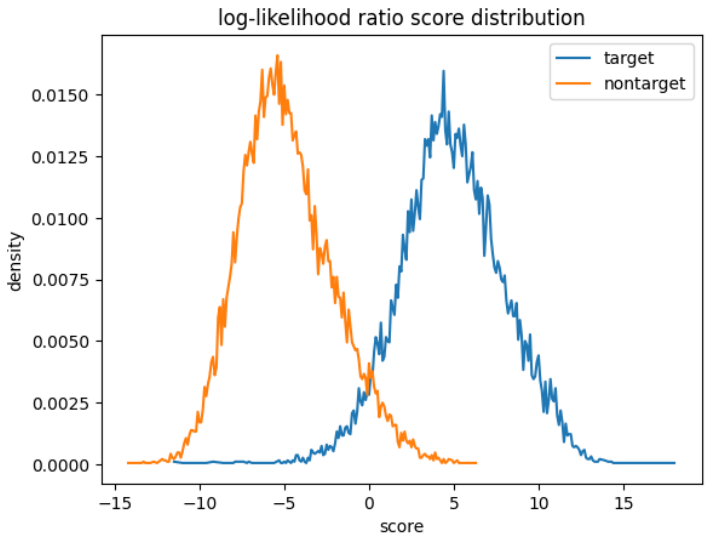
CN-Celeb





VoxCeleb





QMF (Quality Measure Function)

QMF is used to measure the speaker identity-independent information difference between enroll and test.

Calibration with **QMFs (Quality Measure Functions)**: $y = l(s) = w_0 s + w_1 x_1 + w_2 x_2 + ... + w_n x_n + b$

其中:

- y 为 1 (target) or 0 (nontarget)
- s 为 enroll 与 test 的 similarity score
- $l(s)$ 为 s 对应的 log-likelihood ratio
- x_i 为 enroll 与 test 的 quality measure，例如 embedding length ratio、acoustic distance 等
- 学习参数：权重 w_0，w_1，w_2 与偏置 b

$Q(q_e, q_t)$ q_e : quality measure of enroll, q_t : quality measure of test

- $Q = w * \frac{q_e}{q_t}$
- $Q = w * \left| log \frac{q_e}{q_t} \right|$
- $Q = w * log^2 \frac{q_e}{q_t}$

Experiments on CN-Celeb

- Pretrain model: Training on 200 speaker sampling from CN-Celeb

Calibration \ CNC1.Eval	EER (%)	minDCF (0.01)	minDCF (0.001)
non (similarity score)	14.785	0.71413	0.80018
non (similarity score after AS-norm)	14.148	0.71928	0.86103
magnitude	13.861	0.72208	0.86935
duration	13.495	0.72189	0.86616
mean imposter score	13.534	0.70337	**0.85751**
duration + imposter	13.050	0.70520	0.86345
magnitude + duration + imposter	**13.027**	**0.70256**	0.86161

- magnitude, duration QMF can decrease EER, but will increase minDCF
- These three QMFs are somewhat complementary

Experiments on CN-Celeb

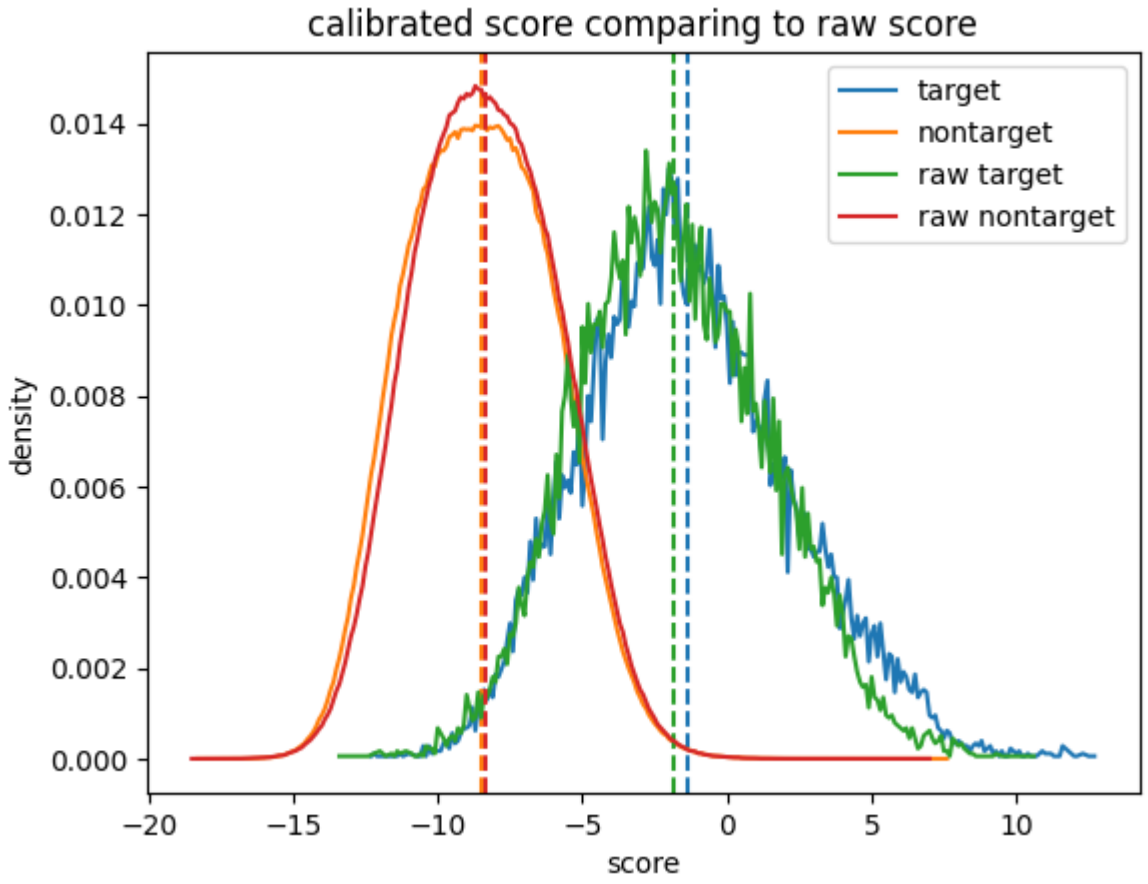
- Pretrain model: Training on 200 speaker sampling from CN-Celeb
- Scoring with multi-enroll embedding average

Calibration \ CNC1.Eval	EER (%)	minDCF (0.01)	minDCF (0.001)
non (similarity score)	14.002	0.70074	0.81121
non (similarity score after AS-norm)	13.309	0.70554	0.84984
magnitude	13.675	0.70682	0.84985
duration	12.757	0.70729	0.86269
mean imposter score	12.565	**0.66610**	**0.83116**
duration + imposter	12.278	0.66824	0.83542
magnitude + duration + imposter	**12.199**	0.66683	0.83499

- Imposter QMF can bring about a big performance improvement, especially minDCF

Score Distribution

CN-Celeb EER: 14.002% --> 12.199%



Experiments on VoxCeleb

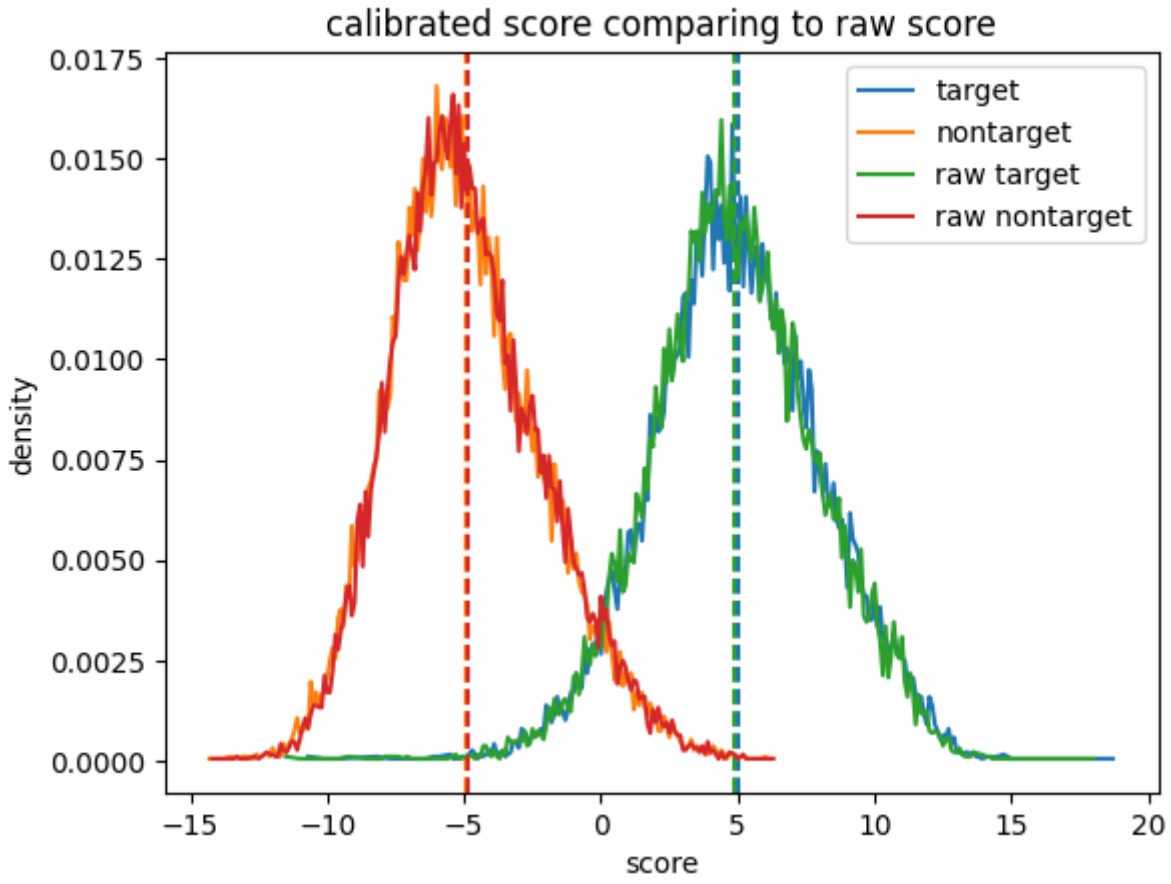
- Pretrain model: Training on 200 speaker sampling from VoxCeleb

Calibration \ Vox1-Clean	EER (%)	minDCF (0.01)	minDCF (0.001)
non (similarity score)	5.611	0.51322	0.65550
non (similarity score after AS-norm)	5.340	0.56127	0.68971
magnitude	5.321	0.56488	0.68897
duration	5.340	0.56368	0.69455
mean imposter score	5.340	0.56127	0.68971
duration + imposter	**5.167**	**0.54229**	0.66541
magnitude + duration + imposter	5.172	0.54868	**0.66504**

- The complementarity of the three QMFs is limited, especially the magnitude

Score Distribution

VoxCeleb EER: 5.611% --> 5.172%



Experiments on **acoustic distance**

- Pretrain model: Training on 200 speaker sampling from CN-Celeb

model \ CNC1.Eval	EER (%)	minDCF (0.01)	minDCF (0.001)
non (similarity score after AS-norm)	14.148	0.71928	0.86103
vggish	14.103	0.72282	0.86547
ECAPA-TDNN_softmax	14.137	0.71661	0.85624
ResNet34L_softmax	14.109	0.71693	0.85752