Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

Interspeech 2024
A4-O2.3 #442

Revisiting and Improving Scoring Fusion for Spoofing-aware Speaker Verification Using Compositional Data Analysis

Xin Wang, Tomi Kinnunen, Kong Aik Lee, Paul-Gauthier Noe, Junichi Yamagishi

NII, JST PRESTO, UEF, PolyU, Inria

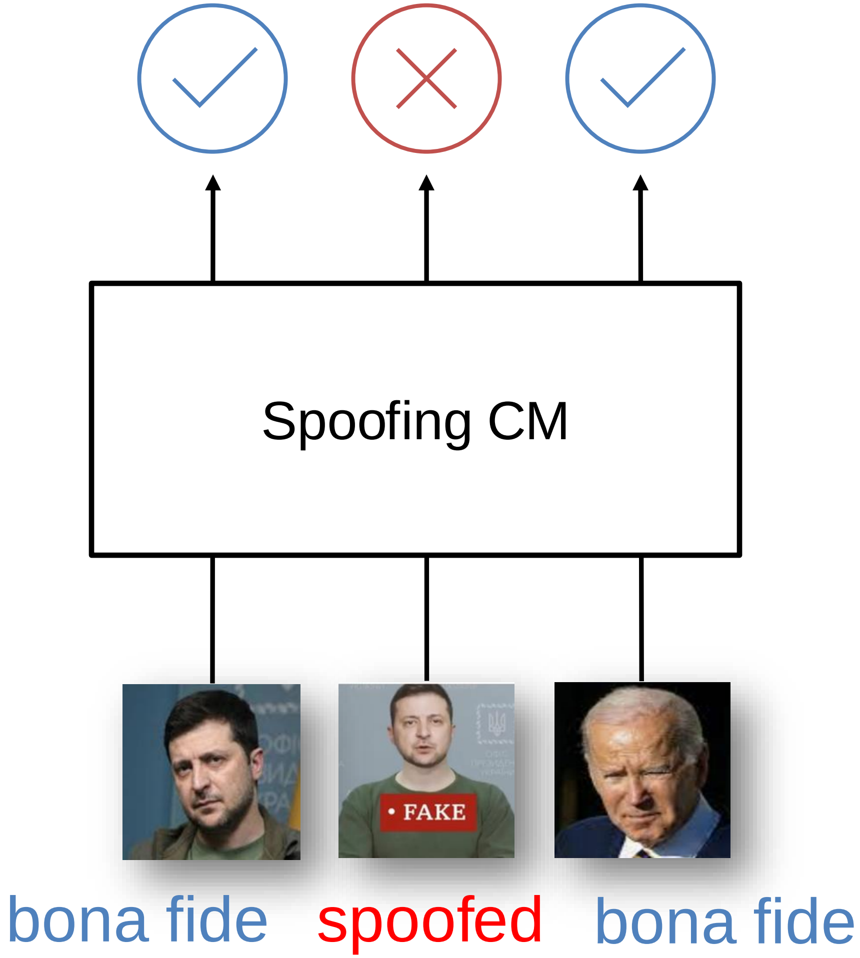
wangxin@nii.ac.jp

Summary in one slide

- Question: how ASV and spoofing countermeasure (CM) should be fused **theoretically**?
- Message: fusing ASV and CM != fusing ASVs (or CMs)
- Methods
  - Linear fusion of log likelihood ratios (LLRs)
  - Non-linear fusion of LLRs

  } Bayesian decision theory
- Results: both better than baseline, non-linear the best

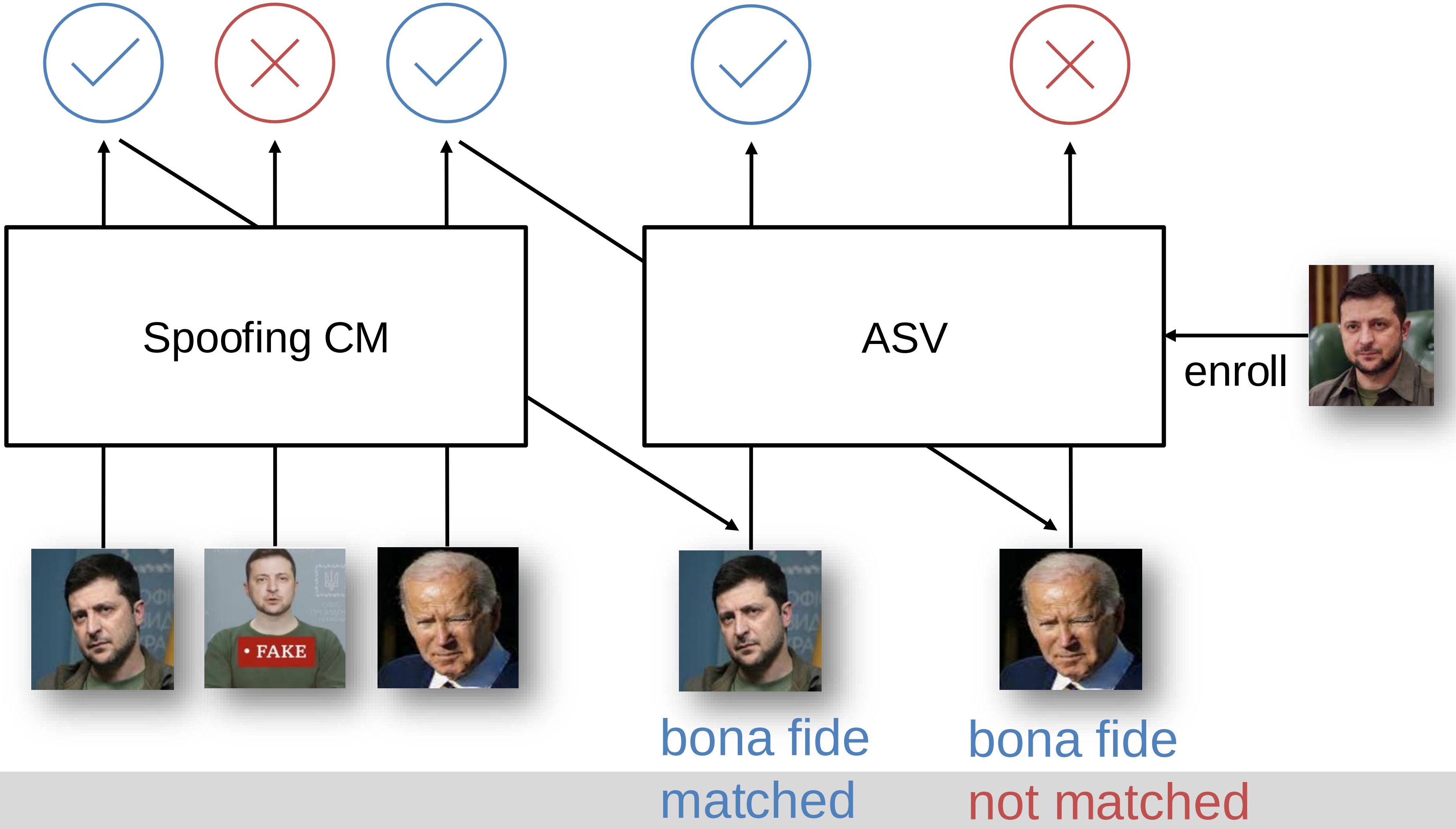
Background: spoofing CM



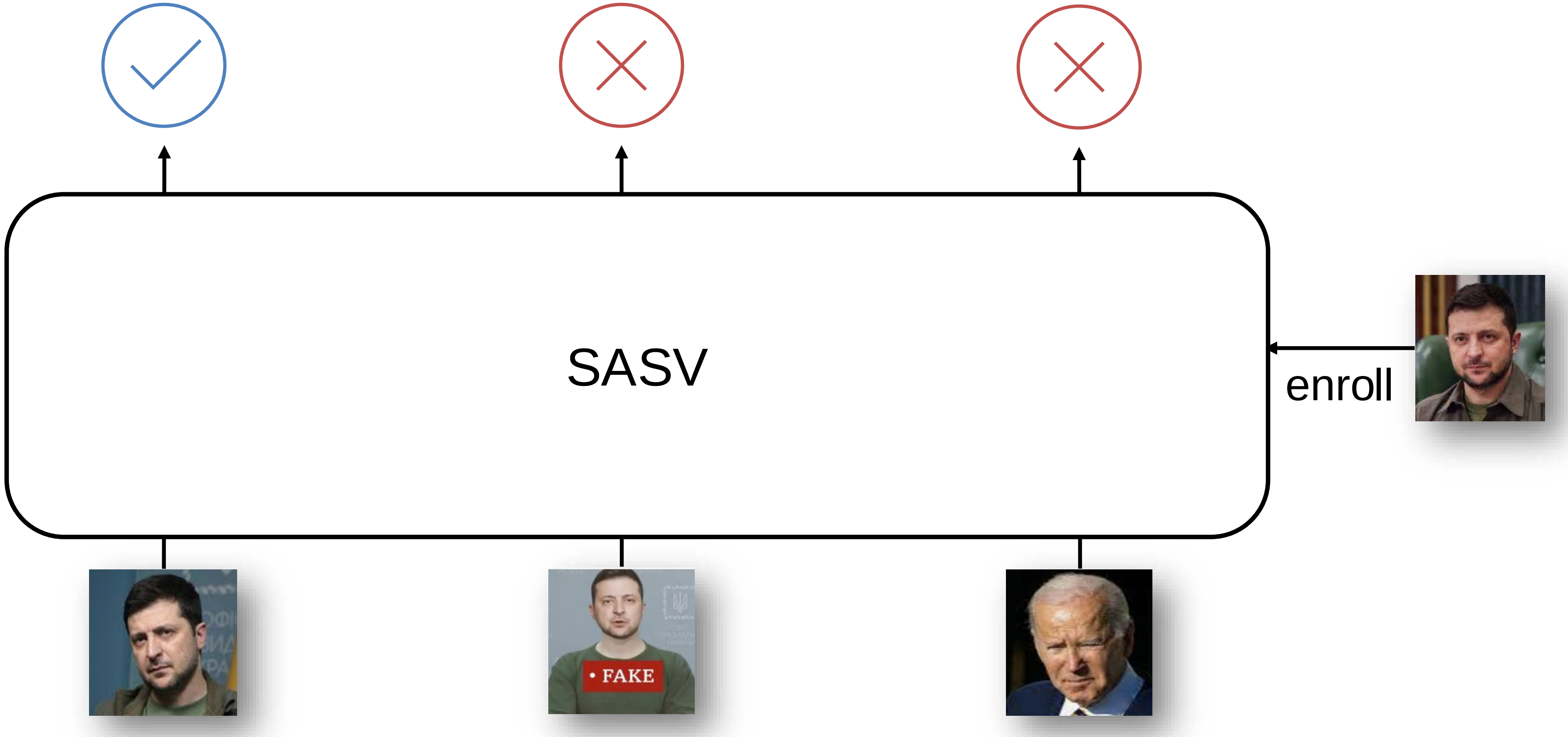
protect human listeners

protect ASV

Background: spoofing CM protecting ASV



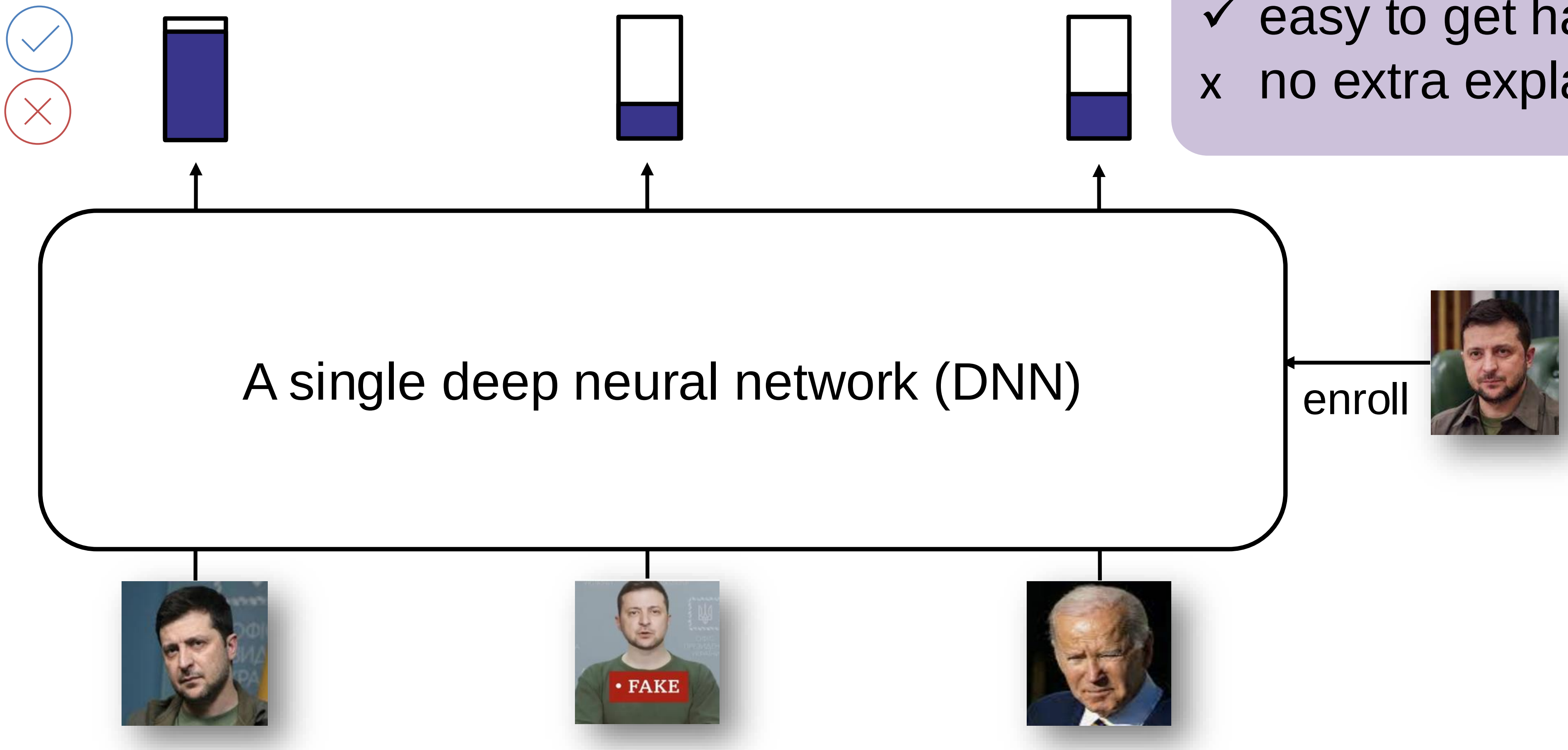
Background: spoofing-robust ASV (SASV)



Background: spoofing-robust ASV (SASV)

❑ Approach 1: end-to-end

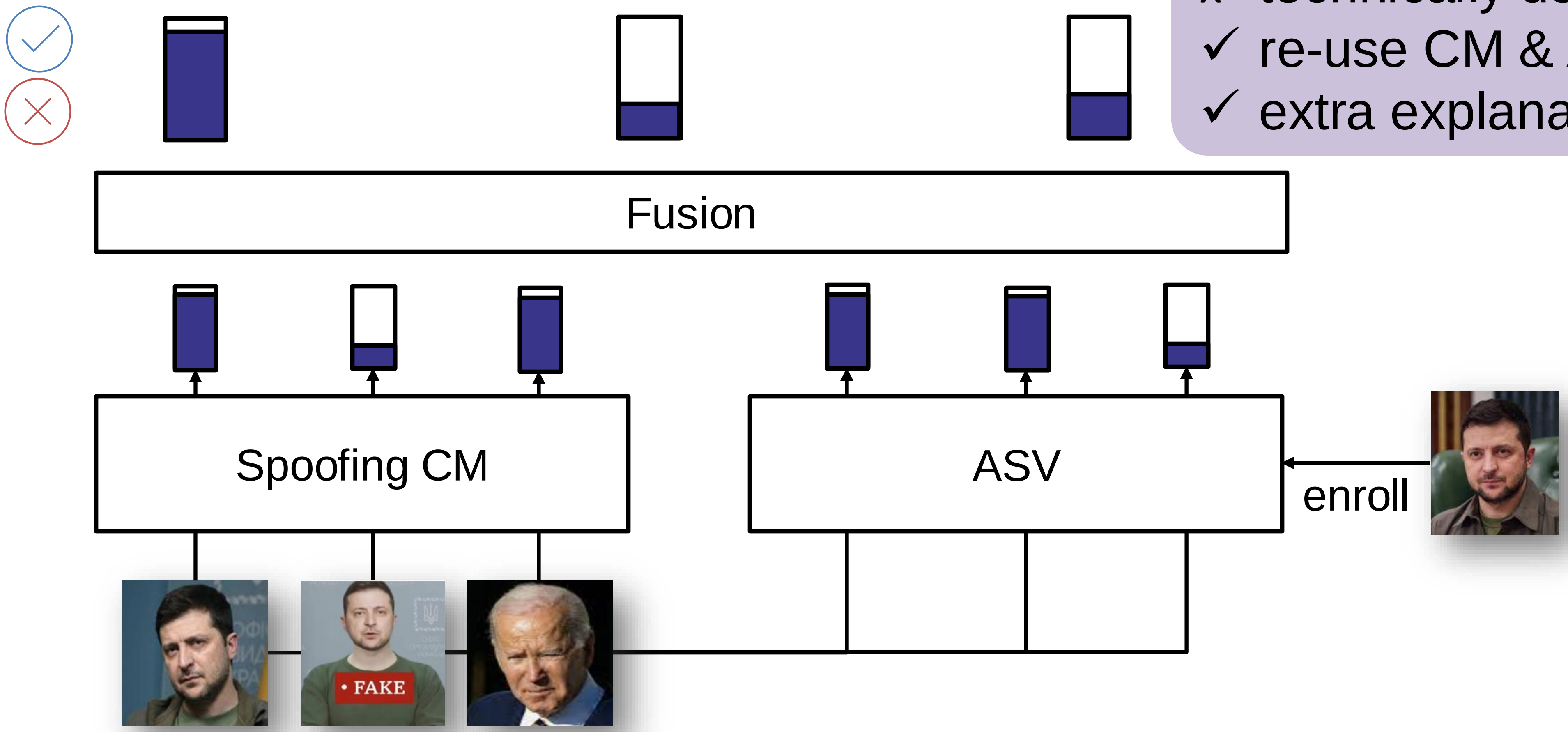
- ✓ easy to get hands on
- x no extra explanation



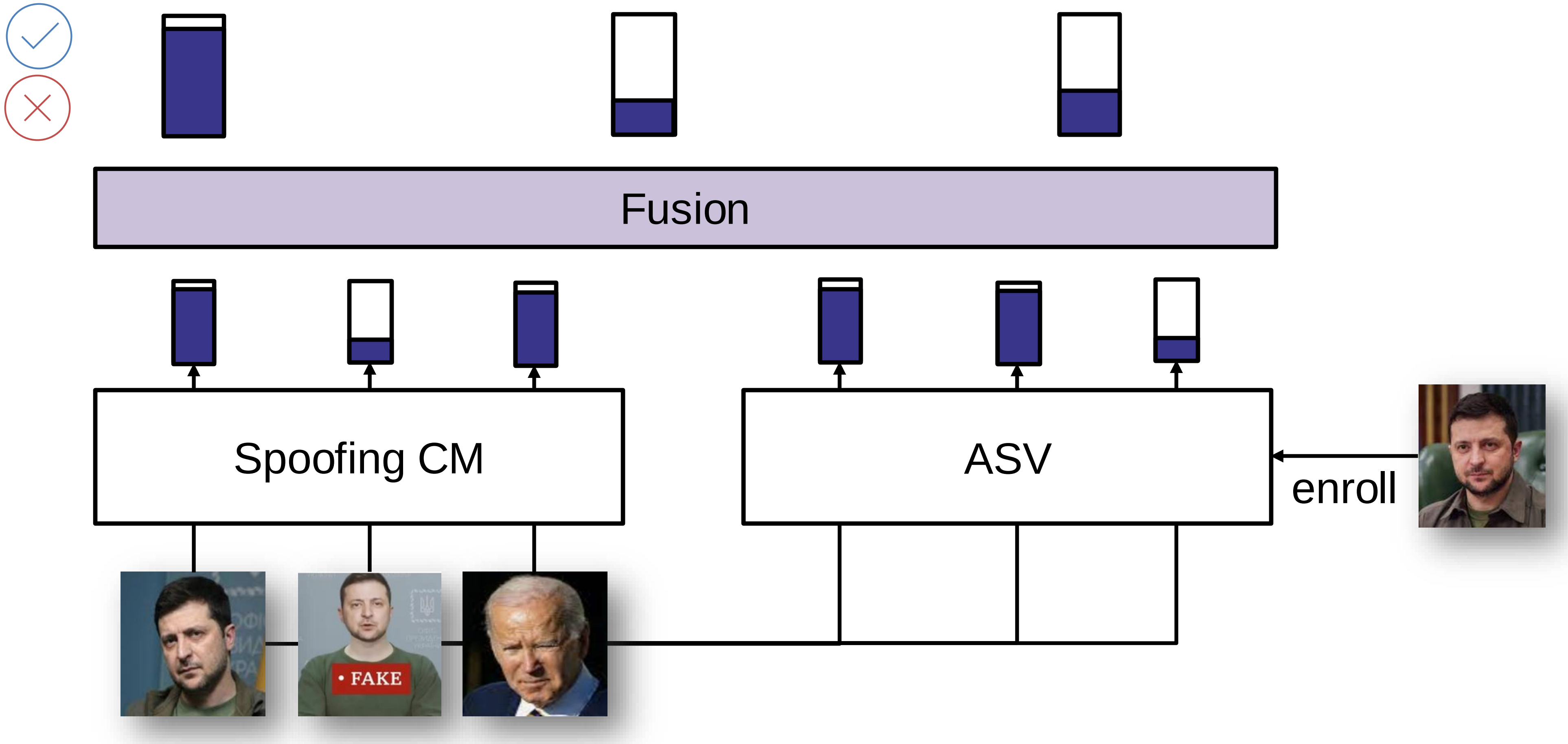
Background: spoofing-robust ASV (SASV)

❑ Approach 2: fusion-based

- x technically demanding
- ✓ re-use CM & ASV
- ✓ extra explanations

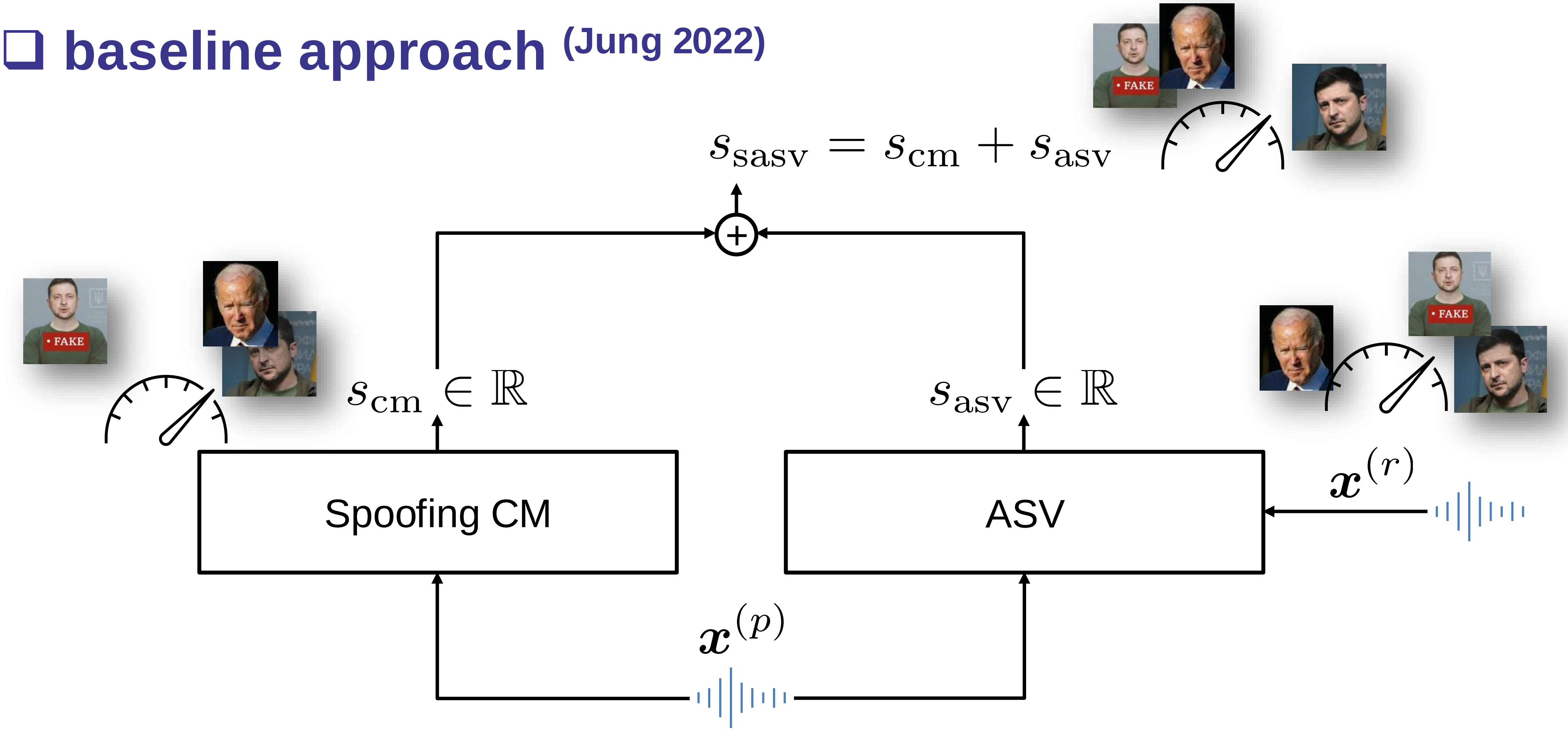


Question: how to *properly* fuse ASV and CM



Question: how to *properly* fuse ASV and CM

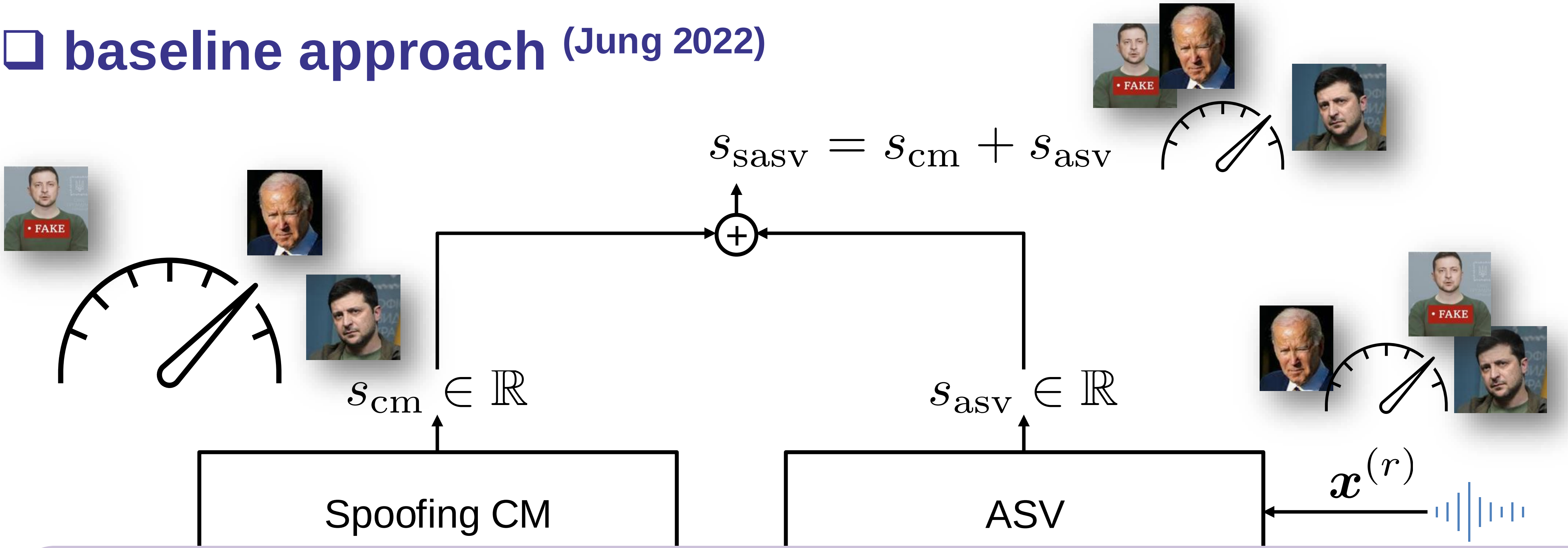
- baseline approach (Jung 2022)



Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. 2022. SASV 2022: The first spoofing-aware speaker verification challenge. In *Proc. Interspeech*, 2022. 2893–2897.

Question: how to *properly* fuse ASV and CM

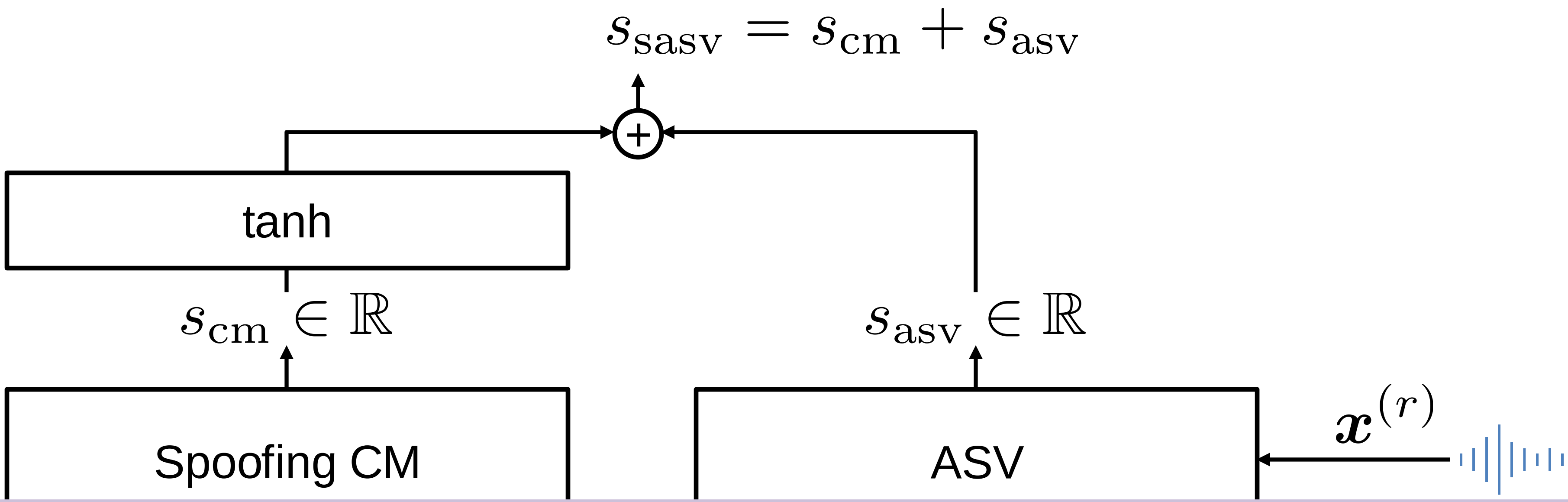
❑ baseline approach (Jung 2022)



? What to do if, say, $s_{\text{cm}} \in [-100, 100]$ $s_{\text{asv}} \in [-1, 1]$

Question: how to *properly* fuse ASV and CM

- **baseline approach** (Jung 2022)



? What to do if, say, $s_{\text{cm}} \in [-100, 100]$ $s_{\text{asv}} \in [-1, 1]$

? Why not normalize both, why summation …

Any thoery to support the good pratice?

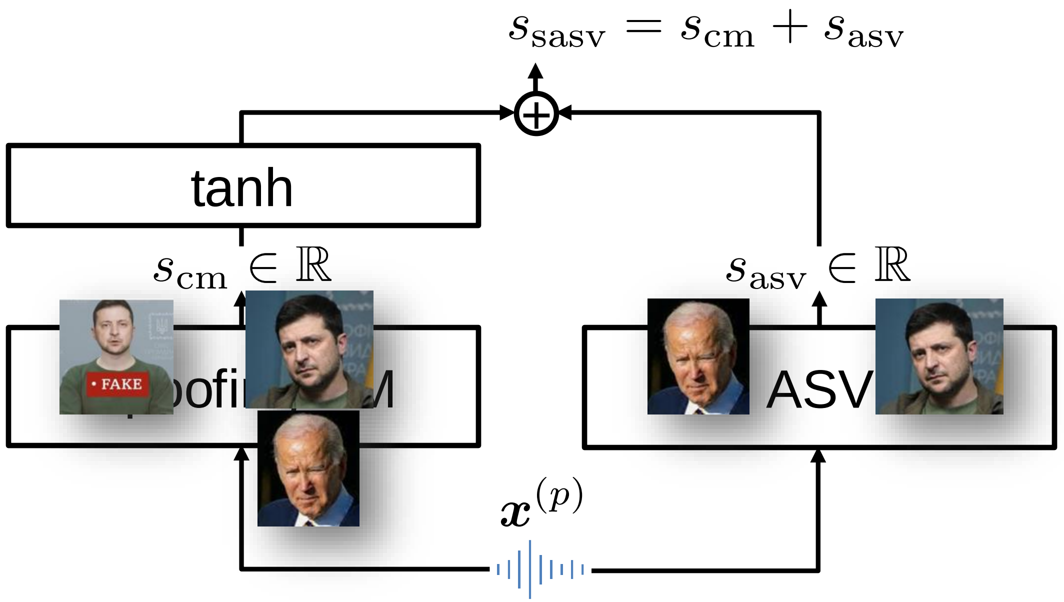
Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. 2022. SASV 2022: The first spoofing-aware speaker verification challenge. In *Proc. Interspeech*, 2022. 2893–2897.

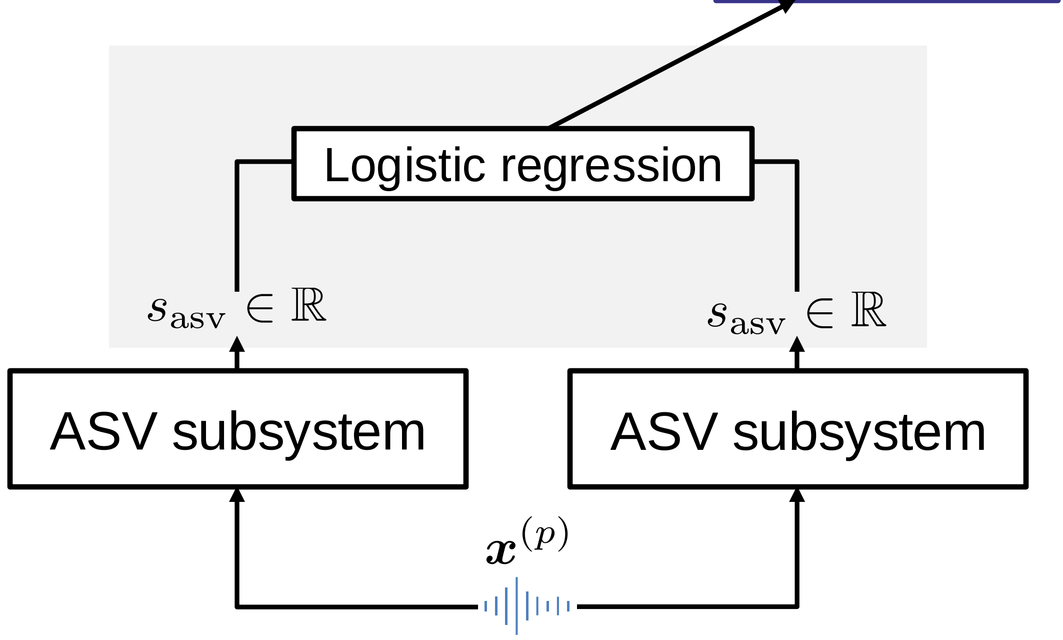
Answers by this work

❑ Fusion in SASV != fusion in ASV (or CM) ensemble (sec.2.1)

- Spoofing CM and ASV are dealing with different pairs of hypotheses
- A different theory is needed

$$\log \frac{P(H_{\text{tar}}|\mathbf{X})}{1 - P(H_{\text{tar}}|\mathbf{X})} = \log \frac{\pi_{\text{tar}}}{1 - \pi_{\text{tar}}} + \sum_{k=1}^{K} \text{llr}_{\text{non}}^{\text{tar}}(\boldsymbol{x}_k),$$





Answers by this work

- **Fusion in SASV != fusion in ASV (or CM) ensemble (sec.2.1)**
  - Spoofing CM and ASV are dealing with different pairs of hypotheses
  - A different theory is needed

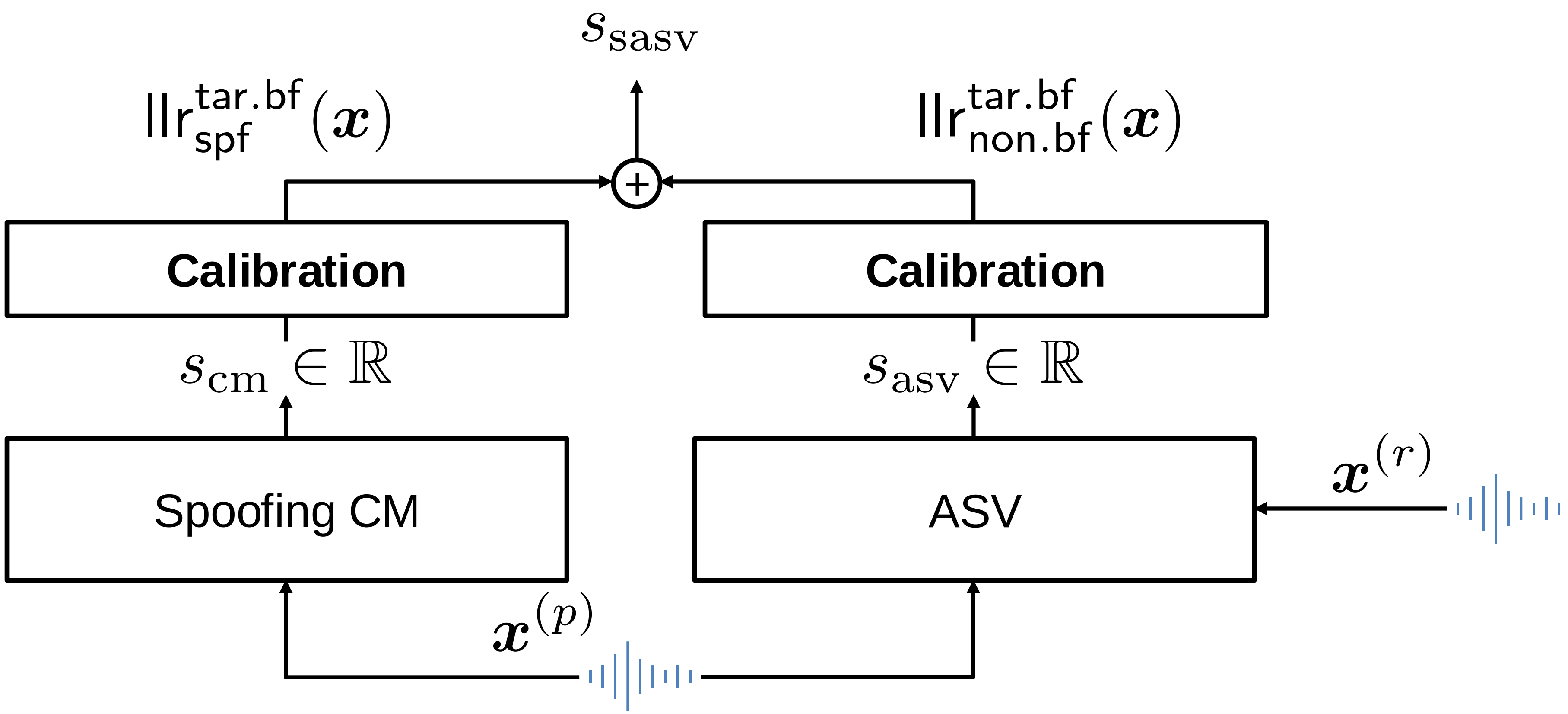
We explain the practice in this talk

- **Linear summation (Sec.2.2 – 2.4)**
  - Bayesian decision theory + compositional data analysis
  - In practice: calibration + sum of CM and ASV LLRs
- **Non-linear fusion (Sec.2.5)**
  - Bayesian decision theory (arxiv appendix)
    - the "optimal" solution to minimize a decision cost
  - In practice: calibration & non-linear fusion

Method 1: linear fusion in good practice

- Score calibrations are needed

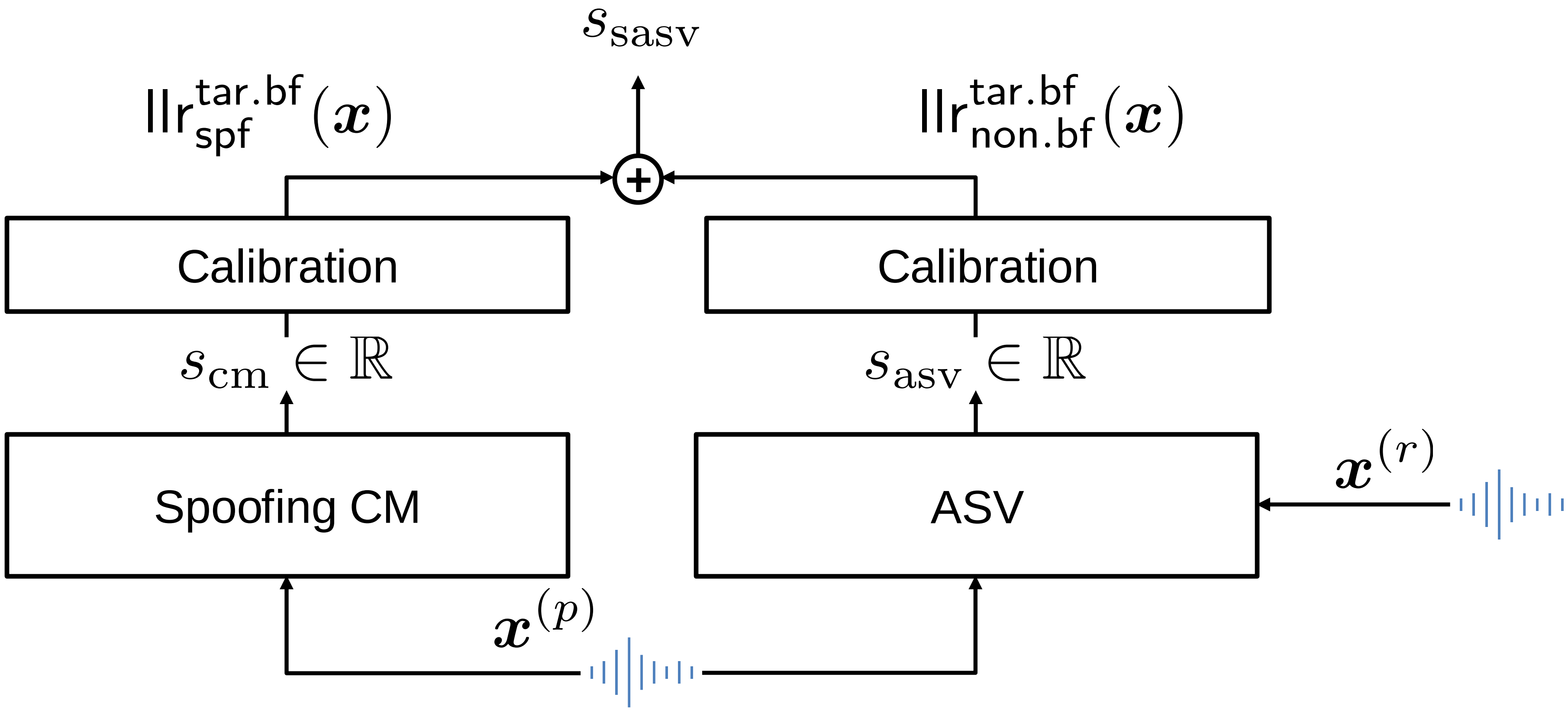
~~? Why normalize s_{cm}, not s_{asv}~~



Method 1: linear fusion in good practice

- Score calibrations are needed
- LLRs should be summed

? ~~Why normalize s_{cm}, not s_{asv}~~
? summation, product

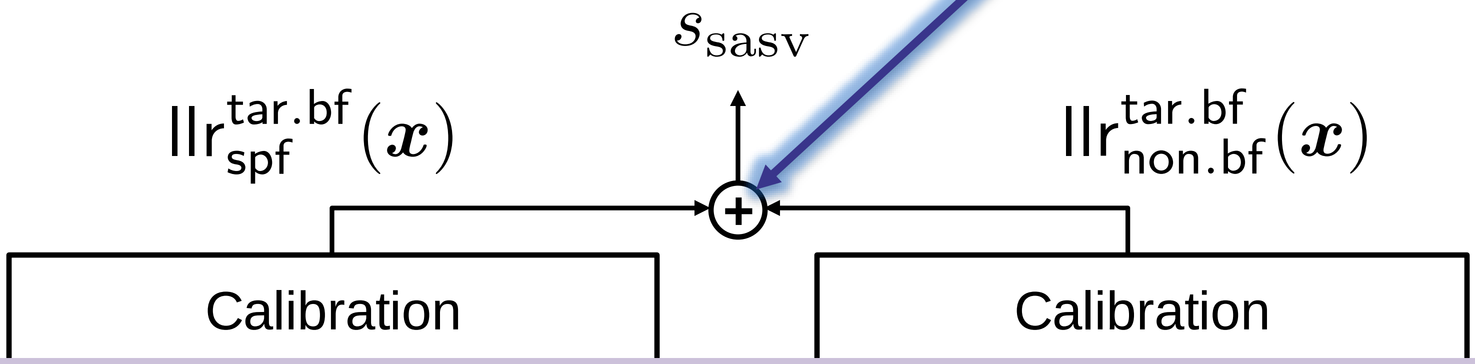


Method 1: linear fusion in good practice

- ❑ **Score calibrations are needed**
- ❑ **LLRs should be summed**

~~? Why normalize s_{cm} , not s_{asv}~~

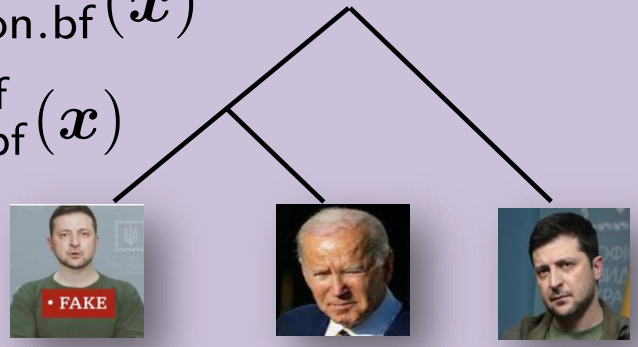
? summation, product



Three data classes but binary decisions!
(sec 2.2 and appendix)

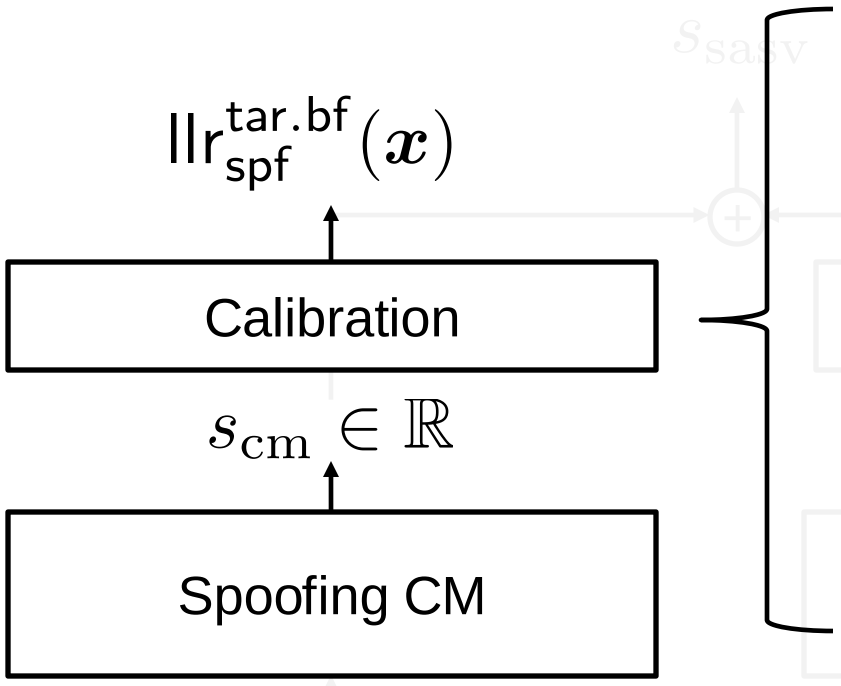
$$s_{\text{sasv}} = \text{llr}_{\text{spf}}^{\text{tar.bf}}(\boldsymbol{x}) + \text{llr}_{\text{non.bf}}^{\text{tar.bf}}(\boldsymbol{x})$$

$$s_{\text{cm}} = \text{llr}_{\text{spf}}^{\text{tar.bf}}(\boldsymbol{x}) - \text{llr}_{\text{non.bf}}^{\text{tar.bf}}(\boldsymbol{x})$$



Method 1: linear fusion in good practice

❑ Score calibration – nothing new



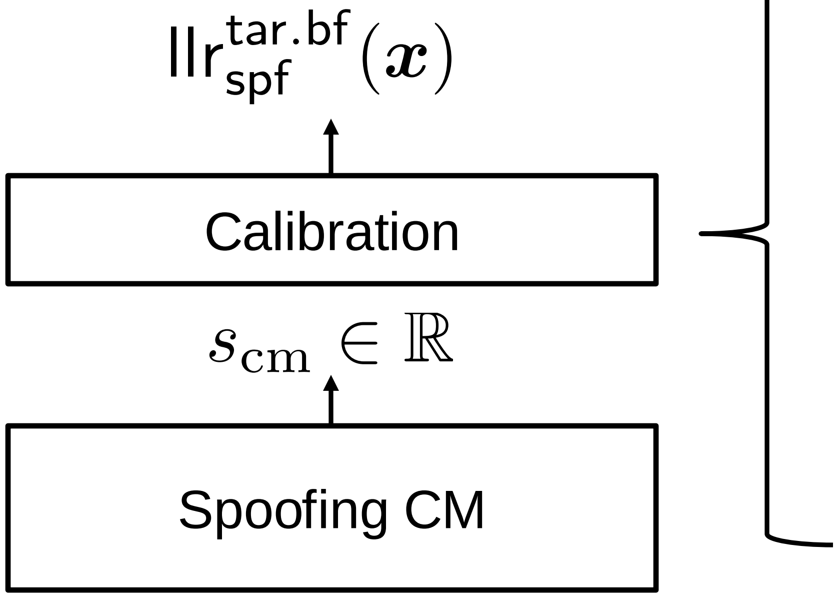
Logistic regression (Morrison 2013)

$$\text{llr}_{\text{non.bf}}^{\text{tar.bf}}(\boldsymbol{x}) = \underline{a} s_{\text{cm}} + \underline{b}$$

estimate {*a*,*b*} on using hold-out data

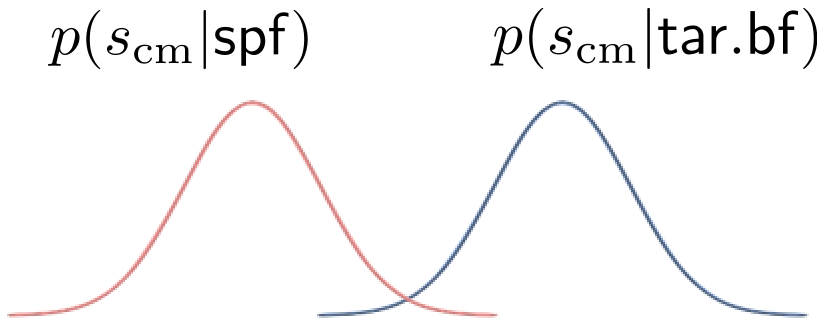
Method 1: linear fusion in good practice

❑ Score calibration – nothing new



Logistic regression

Generative calibration (Brummer 2014)

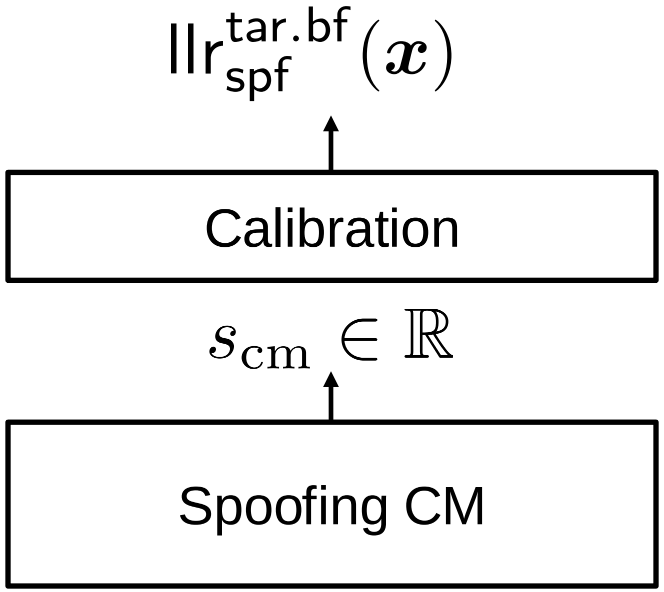


1. choose a parametric distribution
2. estimate distribution para. on dev. set
3. compute $\mathrm{llr}_{\mathrm{spf}}^{\mathrm{tar.bf}}(\boldsymbol{x}) = \log \frac{p(s_{\mathrm{cm}}|\mathrm{tar.bf})}{p(s_{\mathrm{cm}}|\mathrm{spf})}$

Niko Brummer, Albert Swart, and David Van Leeuwen. 2014. A comparison of linear and non-linear calibrations for speaker recognition. In *Proc. Odyssey*, 2014. 14–18.

Method 1: linear fusion in good practice

❑ Score calibration – nothing new



Logistic regression

Generative calibration

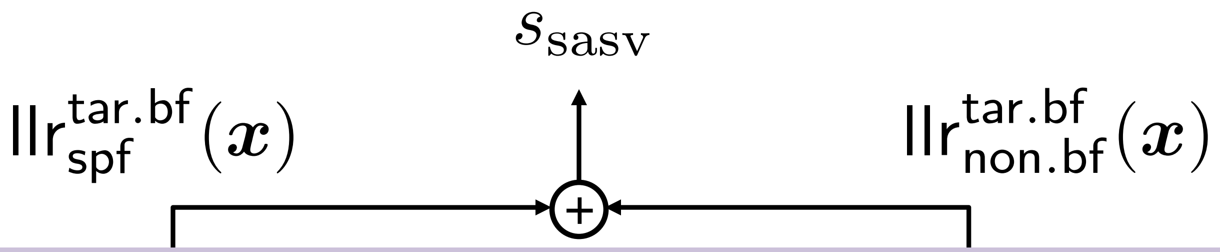
Many other methods exist
(Ferrer 2022, Leeuwen 2013)

Luciana Ferrer, "Analysis and Comparison of Classification Metrics", arXiv:2209.05355, https://github.com/luferrer/CalibrationTutorial
David A. van Leeuwen and Niko Brümmer. 2013. The distribution of calibrated likelihood-ratios in speaker recognition. In *Proc. Interspeech*, 2013. 1619–1623.

Method 1: linear fusion in good practice

❑ Is linear fusion optimal for decision making?

- No



Cost		
Bona fide matched 	0	Cmiss
Bona fide unmatched 	Cfa	0
Spoofed 	Cfa	0

See more in Sec2.5 & Appendix

Method 2: non-linear fusion is better

❑ Non-linear fusion minimizes the cost

$$s_{\text{sasv}} = -\log\left[(1-\rho)e^{-\text{llr}^{\text{tar.bf}}_{\text{non.bf}}} + \rho e^{-\text{llr}^{\text{tar.bf}}_{\text{spf}}}\right]$$

for Cfa=Cmiss

$\text{llr}^{\text{tar.bf}}_{\text{spf}}(\boldsymbol{x})$ → fuse ← $\text{llr}^{\text{tar.bf}}_{\text{non.bf}}(\boldsymbol{x})$

See more in Sec2.5 & Appendix

Cost	✓	✗
Bona fide matched 	0	Cmiss
Bona fide unmatched 	Cfa	0
Spoofed 	Cfa	0

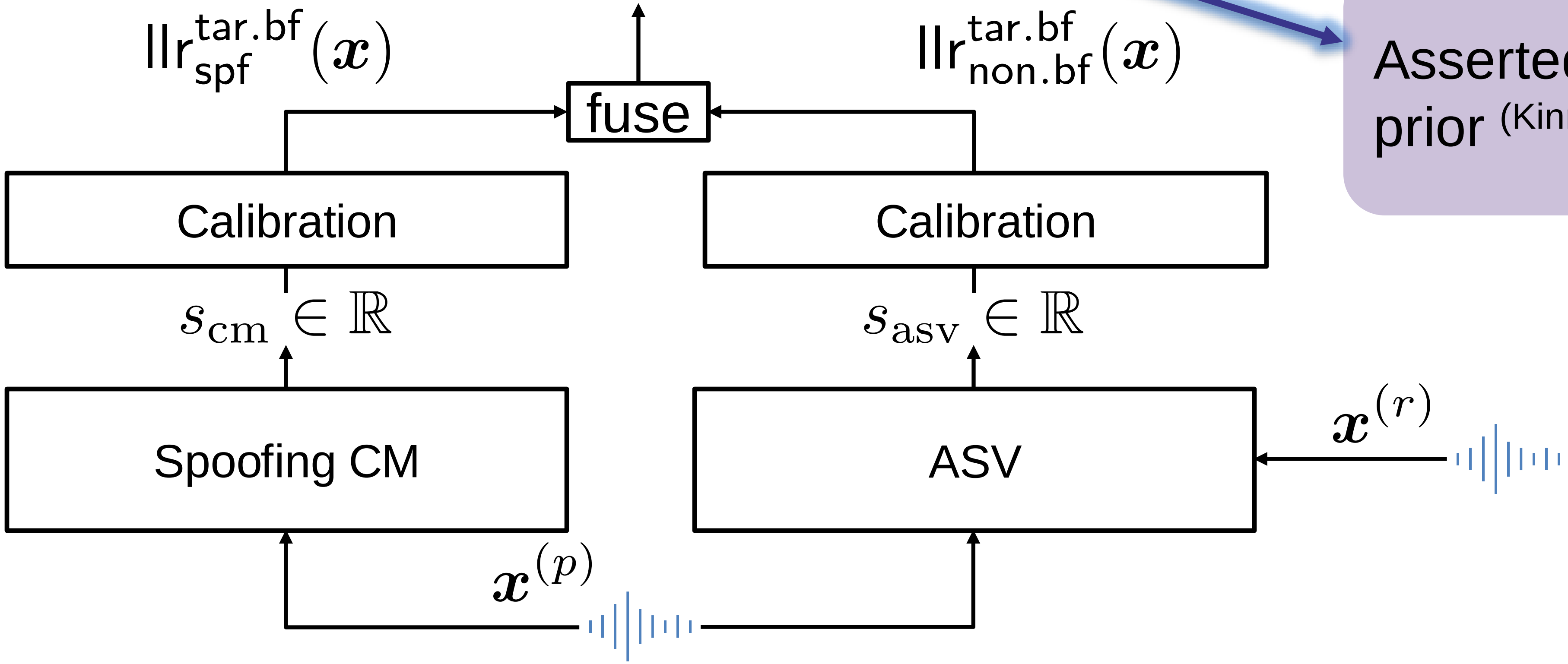
Method 2: non-linear fusion is better

❑ Non-linear fusion minimizes the cost

$$s_{\mathrm{sasv}} = -\log\left[(1-\rho)e^{-\mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{non.bf}}} + \rho e^{-\mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{spf}}}\right]$$

for Cfa=Cmiss

Asserted spoofing prior (Kinnuen 2023)

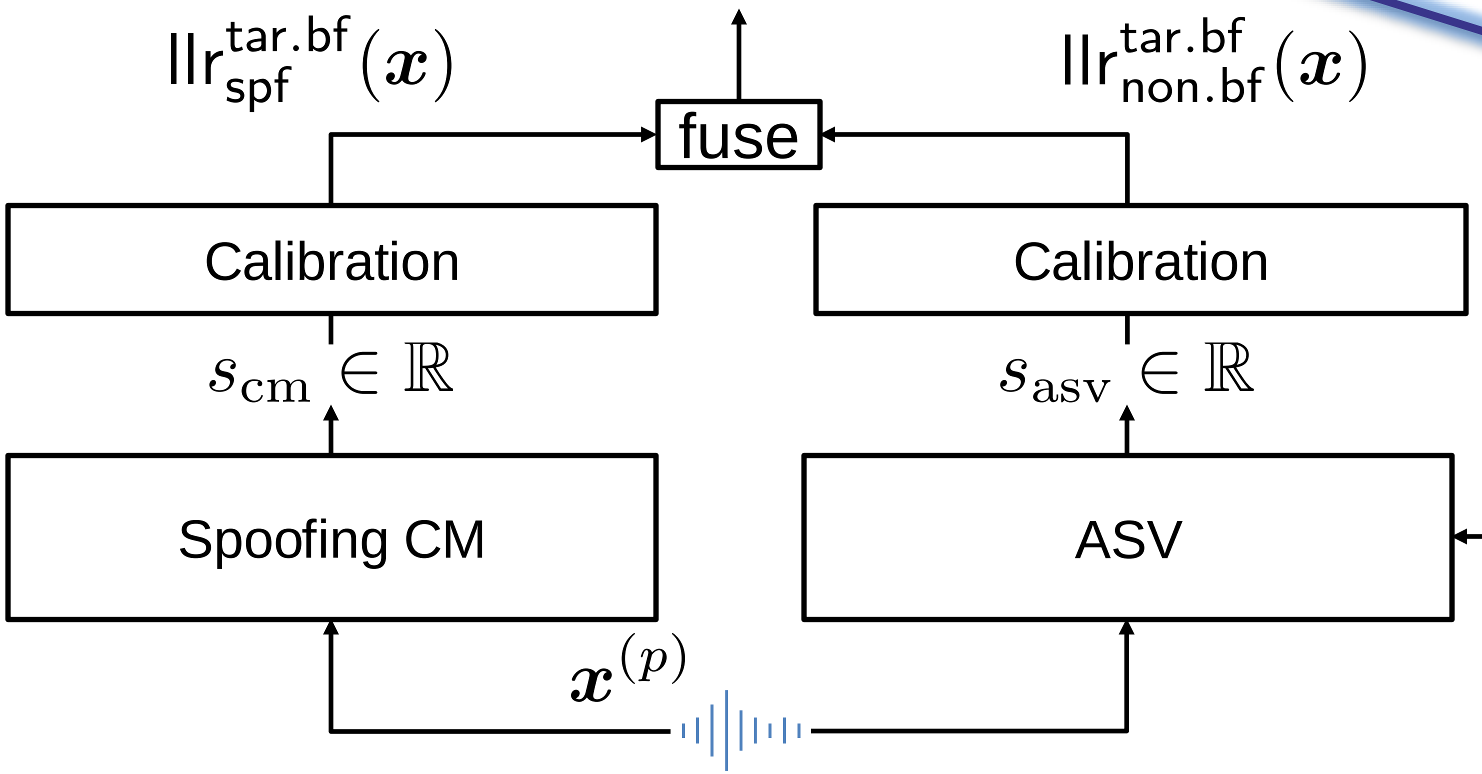


Method 2: non-linear fusion is better

❑ Non-linear fusion minimizes the cost

$$s_{\text{sasv}} = -\log\left[(1-\rho)e^{-\text{llr}^{\text{tar.bf}}_{\text{non.bf}}} + \rho e^{-\text{llr}^{\text{tar.bf}}_{\text{spf}}}\right]$$

for Cfa=Cmiss



Asserted spoofing prior (Kinnuen 2023)

A general form of ASV ($\rho = 0$) or CM ($\rho = 1$)

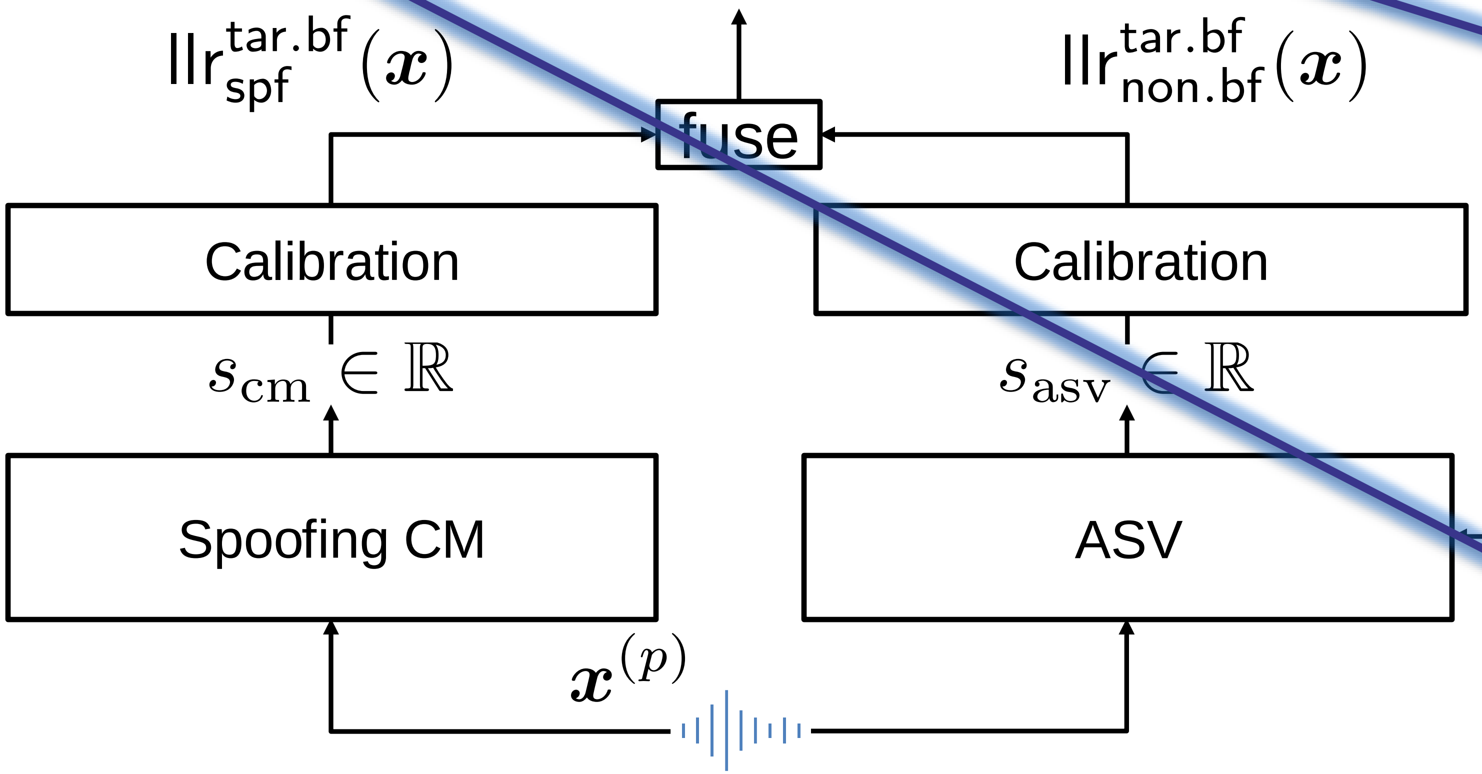
Tomi H. Kinnunen, Kong Aik Lee, Hemlata Tak, Nicholas Evans, and Andreas Nautsch. 2023. t-EER: Parameter-Free Tandem Evaluation of Countermeasures and Biometric Comparators. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023), 1–16. https://doi.org/10.1109/TPAMI.2023.3313648

Method 2: non-linear fusion is better

- Non-linear fusion minimizes the cost

$$s_{\text{sasv}} = -\log\left[(1-\rho)e^{-\text{llr}^{\text{tar.bf}}_{\text{non.bf}}} + \rho e^{-\text{llr}^{\text{tar.bf}}_{\text{spf}}}\right]$$

for Cfa=Cmiss



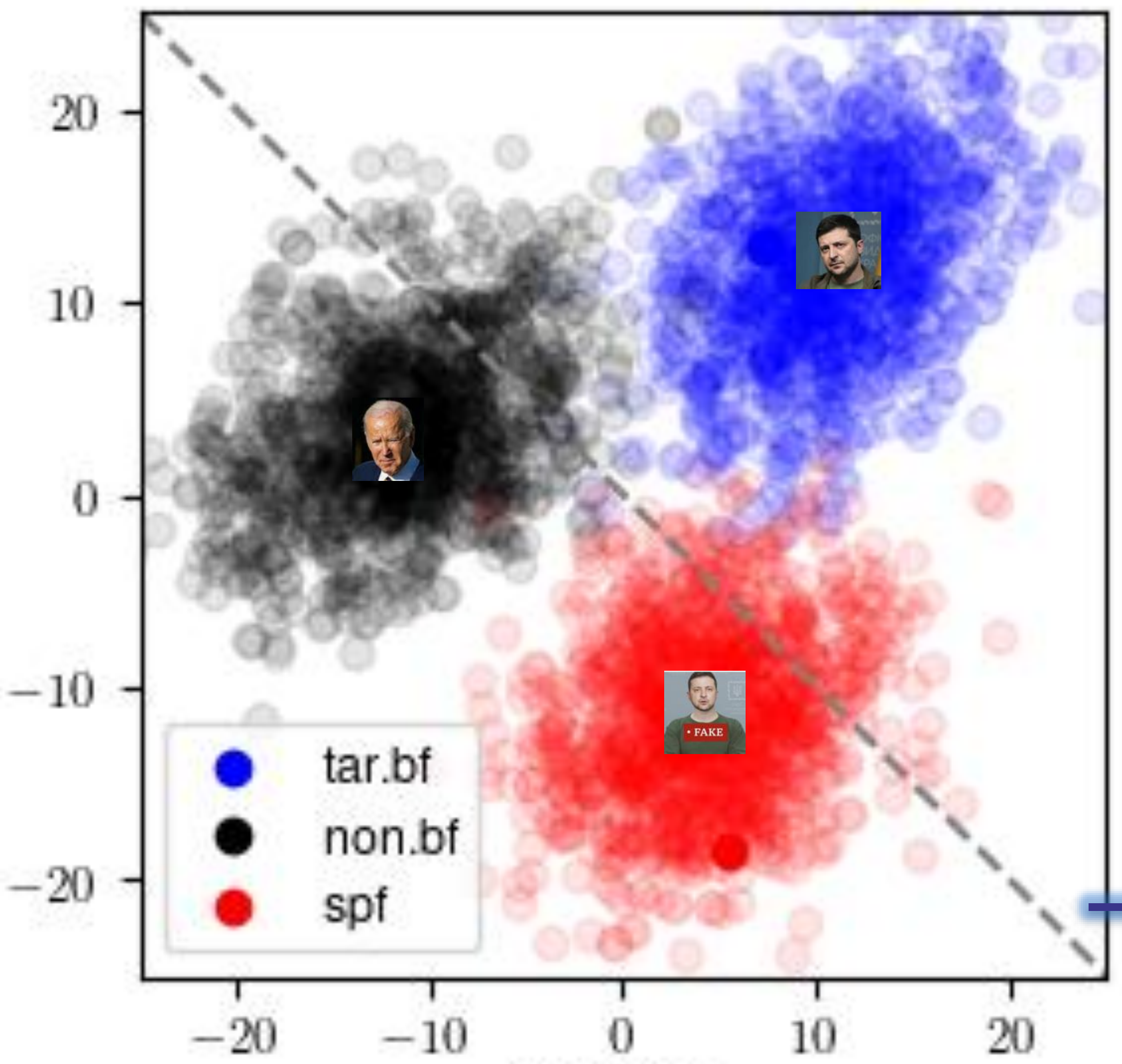
Asserted spoofing prior (Kinnuen 2023)

A general form of ASV ($\rho = 0$) or CM ($\rho = 1$)

A general form of Gaussian fusion (Todisco 2018)

Massimiliano Todisco, Héctor Delgado, Kong Aik Lee, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi. 201… presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion. In Proc. Interspeech, 2018. 77–81.

Demo on toy data set

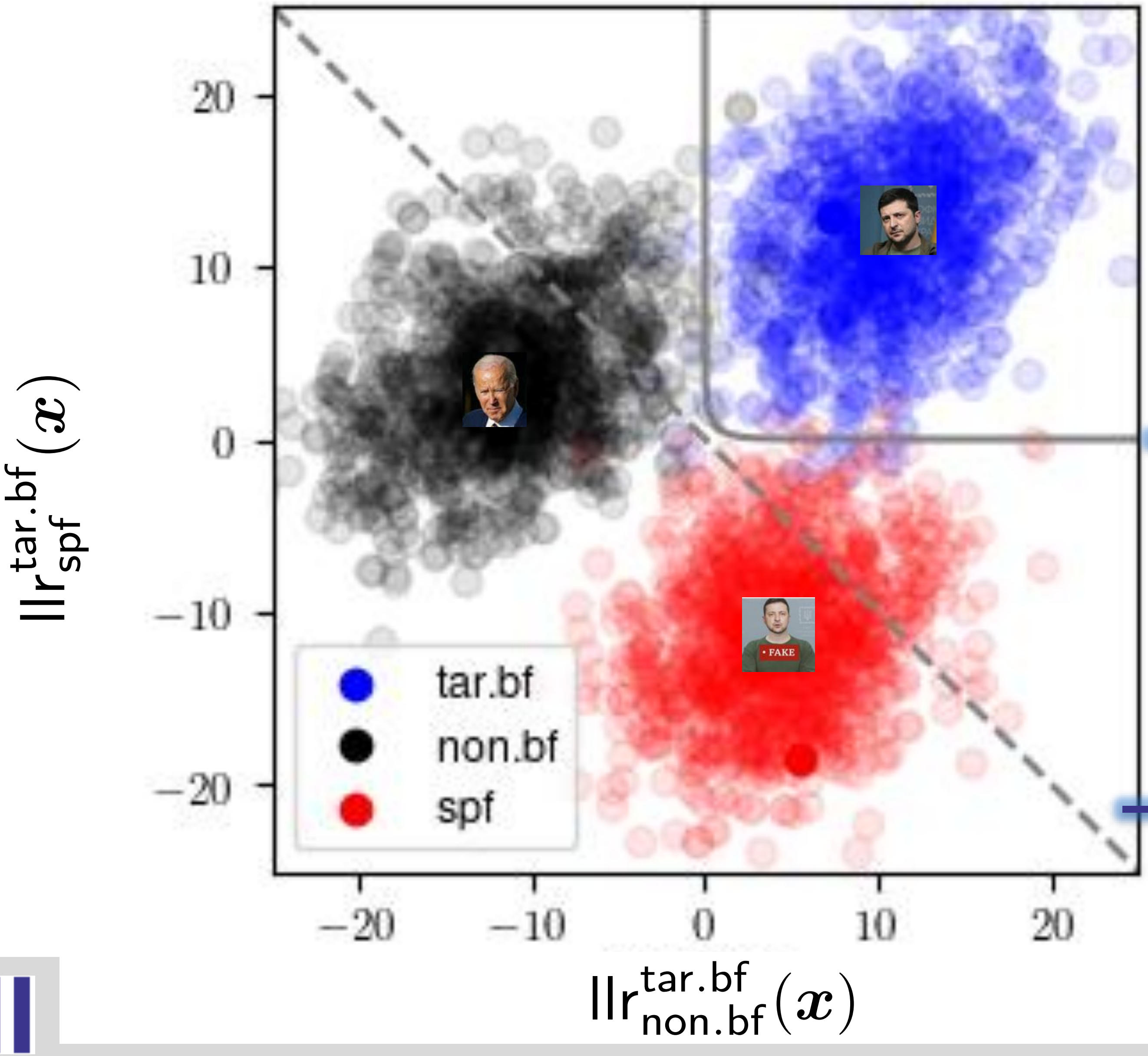


$\mathsf{llr}^{\mathsf{tar.bf}}_{\mathsf{spf}}(\boldsymbol{x})$

$\mathsf{llr}^{\mathsf{tar.bf}}_{\mathsf{non.bf}}(\boldsymbol{x})$

$s_{\text{sasv}} = \mathsf{llr}^{\mathsf{tar.bf}}_{\mathsf{non.bf}} + \mathsf{llr}^{\mathsf{tar.bf}}_{\mathsf{spf}}$

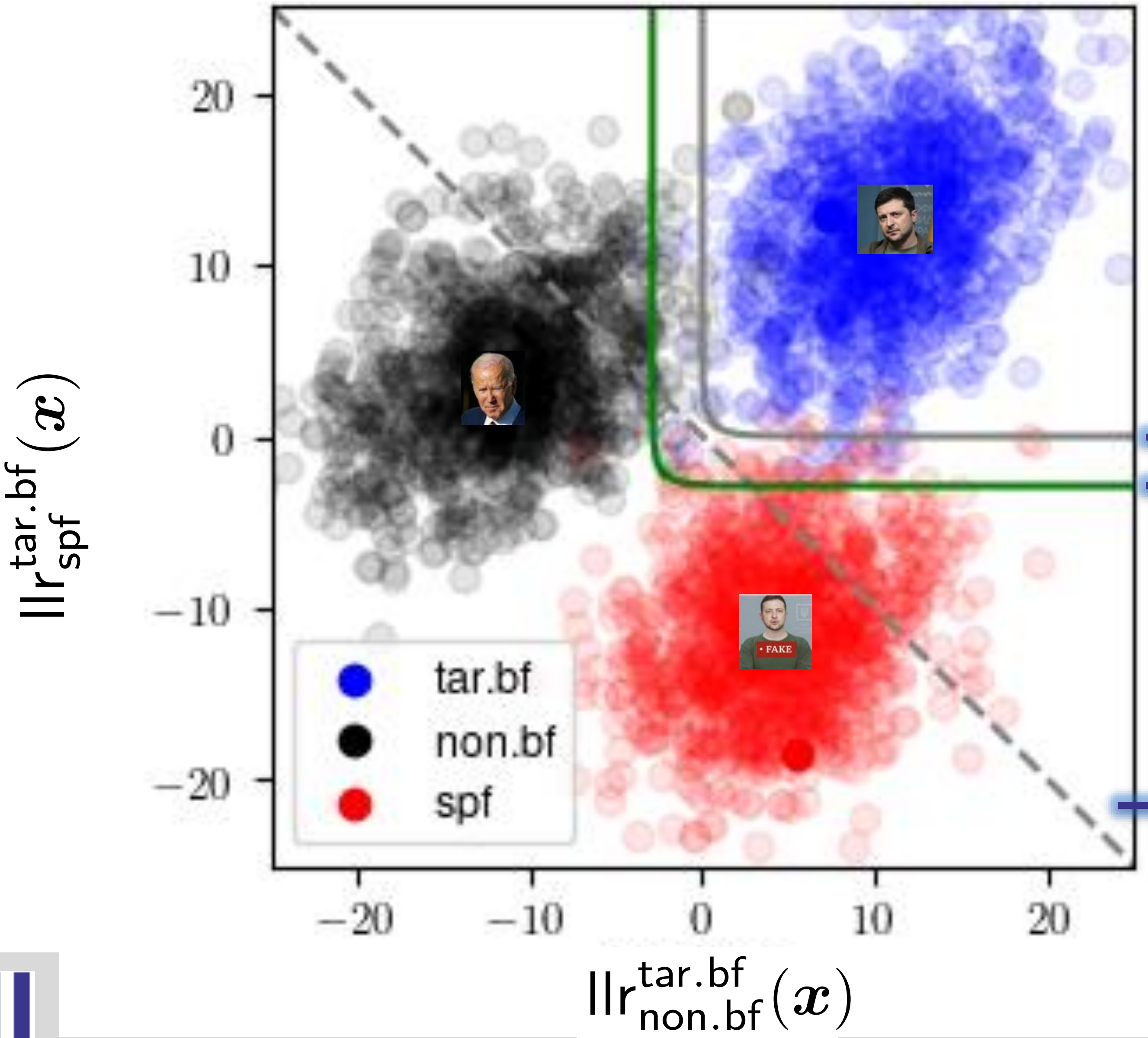
Demo on toy data set



$$s_{\mathrm{sasv}} = -\log\left[(1-\rho)e^{-\mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{non.bf}}} + \rho e^{-\mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{spf}}}\right]$$

$$s_{\mathrm{sasv}} = \mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{non.bf}} + \mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{spf}}$$

Demo on toy data set



$$s_{\mathrm{sasv}} = -\log\left[(1-\rho)e^{-\mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{non.bf}}} + \rho e^{-\mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{spf}}}\right]$$

different ρ

$$s_{\mathrm{sasv}} = \mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{non.bf}} + \mathrm{llr}^{\mathrm{tar.bf}}_{\mathrm{spf}}$$

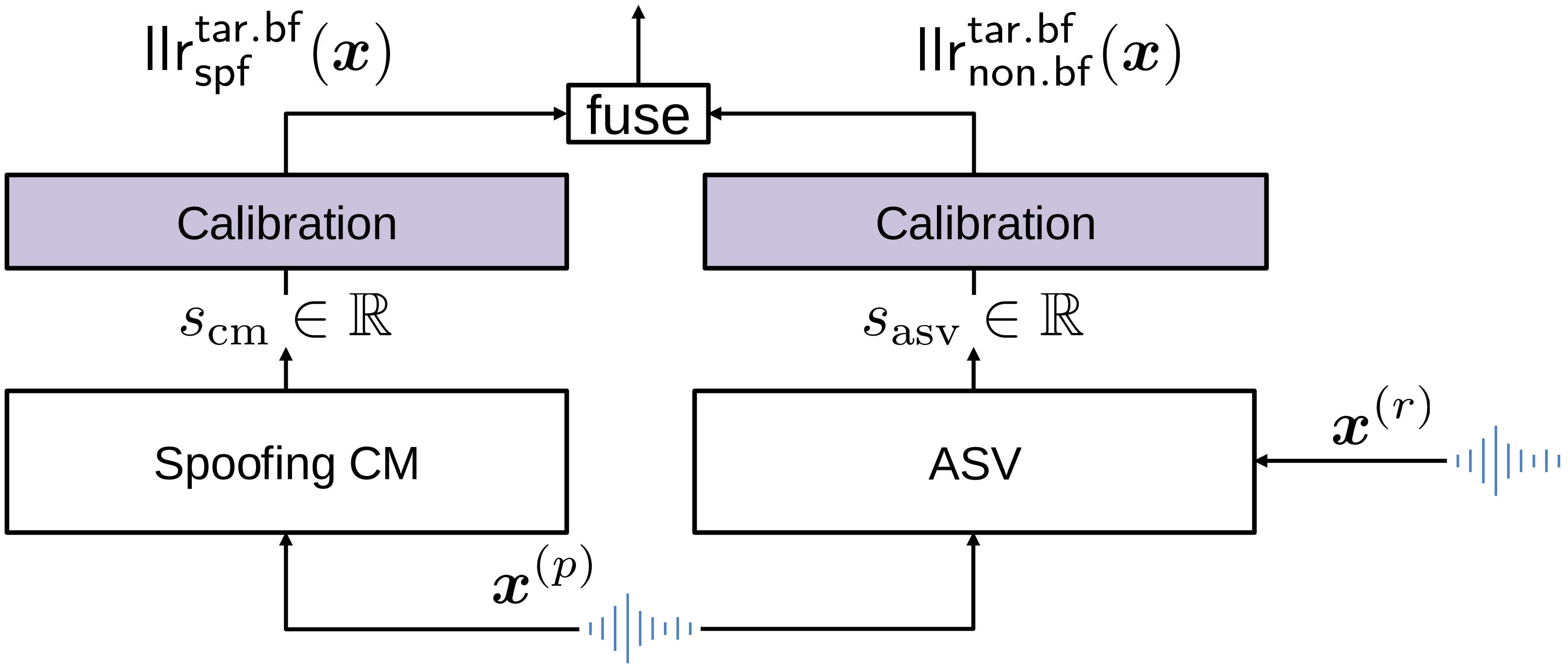
Recap the practices

Linear fusion

$$s_{\text{sasv}} = \text{llr}^{\text{tar.bf}}_{\text{non.bf}} + \text{llr}^{\text{tar.bf}}_{\text{spf}}$$

Non-linear fusion

$$s_{\text{sasv}} = -\log\left[(1-\rho)e^{-\text{llr}^{\text{tar.bf}}_{\text{non.bf}}} + \rho e^{-\text{llr}^{\text{tar.bf}}_{\text{spf}}}\right]$$



All are supported by decision theory

Experiments

❑ Data

- SASV 2022 challenge database, official protocols (Jung 2022)

❑ Systems

- All use *pre-trained* ASV and CM from SASV 2022 B1 (Jung 2022)
- Systems differ in score calibration & fusion

❑ Misc

- Training & evaluation in six rounds
- Averaged results are reported

Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. 2022. SASV 2022: The first spoofing-aware speaker verification challenge. In Proc. Interspeech, 2022. 2893–2897.

Experiments

☹ worse → ☺ better

ID	B1	B1c	L2	L2c	L3	L3c	B1v2	Post
Fusion	linear		linear		non-linear			
Calibration	×	✓	×	✓	×	✓	×	×
SASV-EER (%)	20.46	2.73	3.31	1.56	1.44	1.43	1.60	1.55
conf. ($\alpha = 5\%$)	±0.40	±0.27	±0.31	±0.23	±0.23	±0.23	±0.22	±0.24
Cllr	2.17	1.09	1.04	0.14	0.18	0.16	0.96	0.84
Cllr_{min}	0.52	0.11	0.13	0.07	0.06	0.07	0.08	0.07
$\text{Cllr}_{\text{calib}}$	1.64	0.98	0.91	0.07	0.11	0.10	0.88	0.78
t-EER (%)	2.10	2.10	1.68	1.68	1.68	1.68	2.19	2.21

SASV-EER (Jung2022)

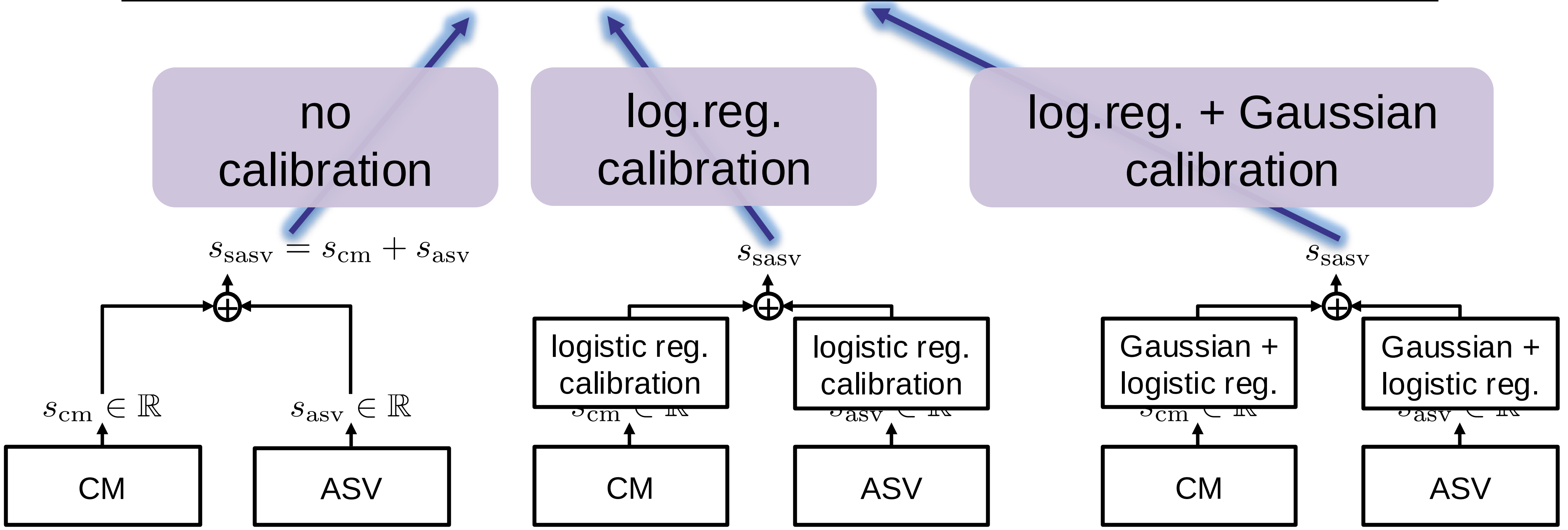
other metrics

Systems with different fusion & calibration methods

From other papers

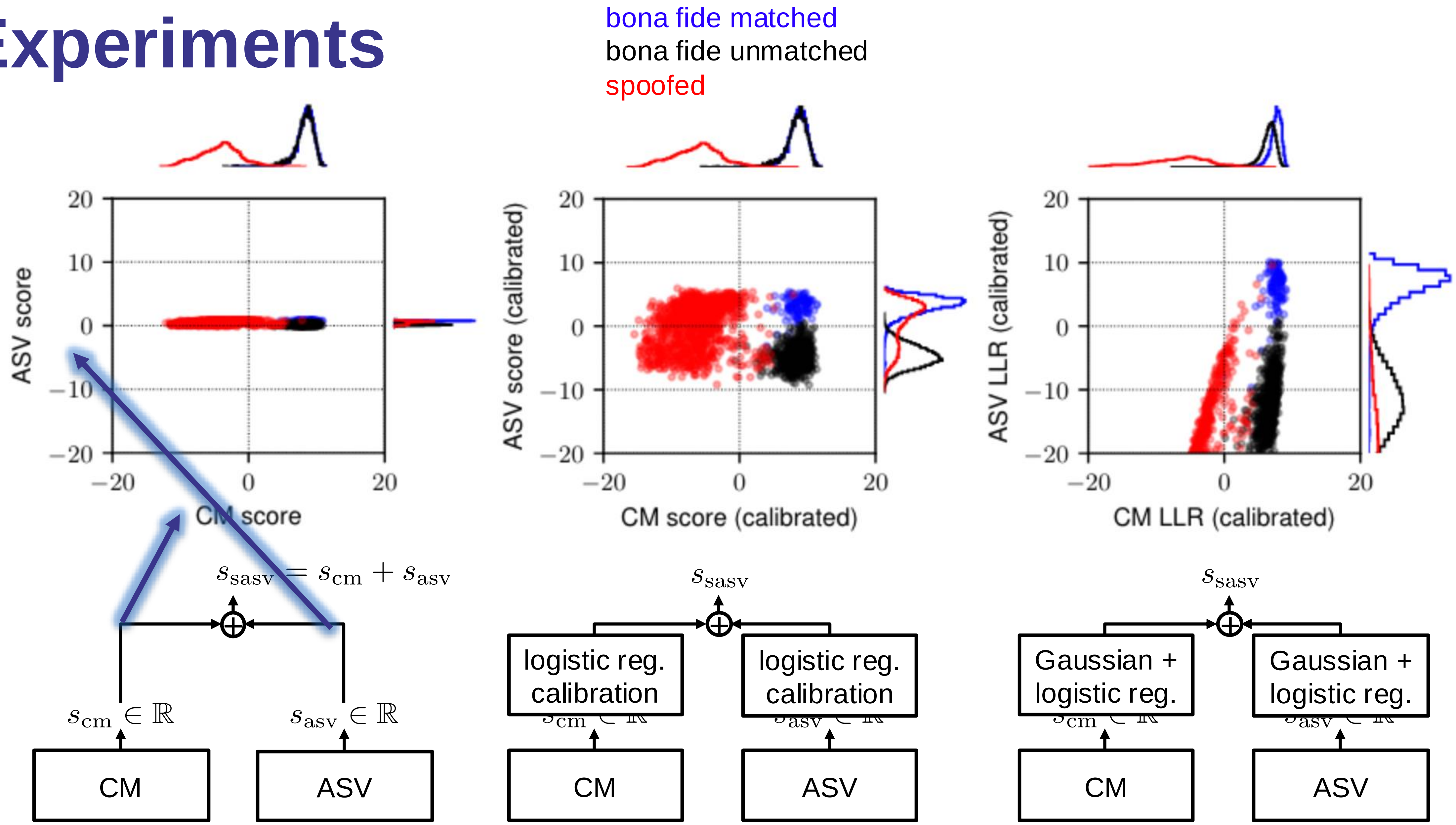
Experiments

ID	B1	B1c	L2	L2c	L3	L3c	B1v2	Post
Fusion	linear			linear				
Calibration	×	✓	×	✓	×	✓	×	×
SASV-EER (%)	20.46	2.73	3.31	1.56	1.44	1.43	1.60	1.55
conf. ($\alpha = 5\%$)	±0.40	±0.27	±0.31	±0.23	±0.23	±0.23	±0.22	±0.24



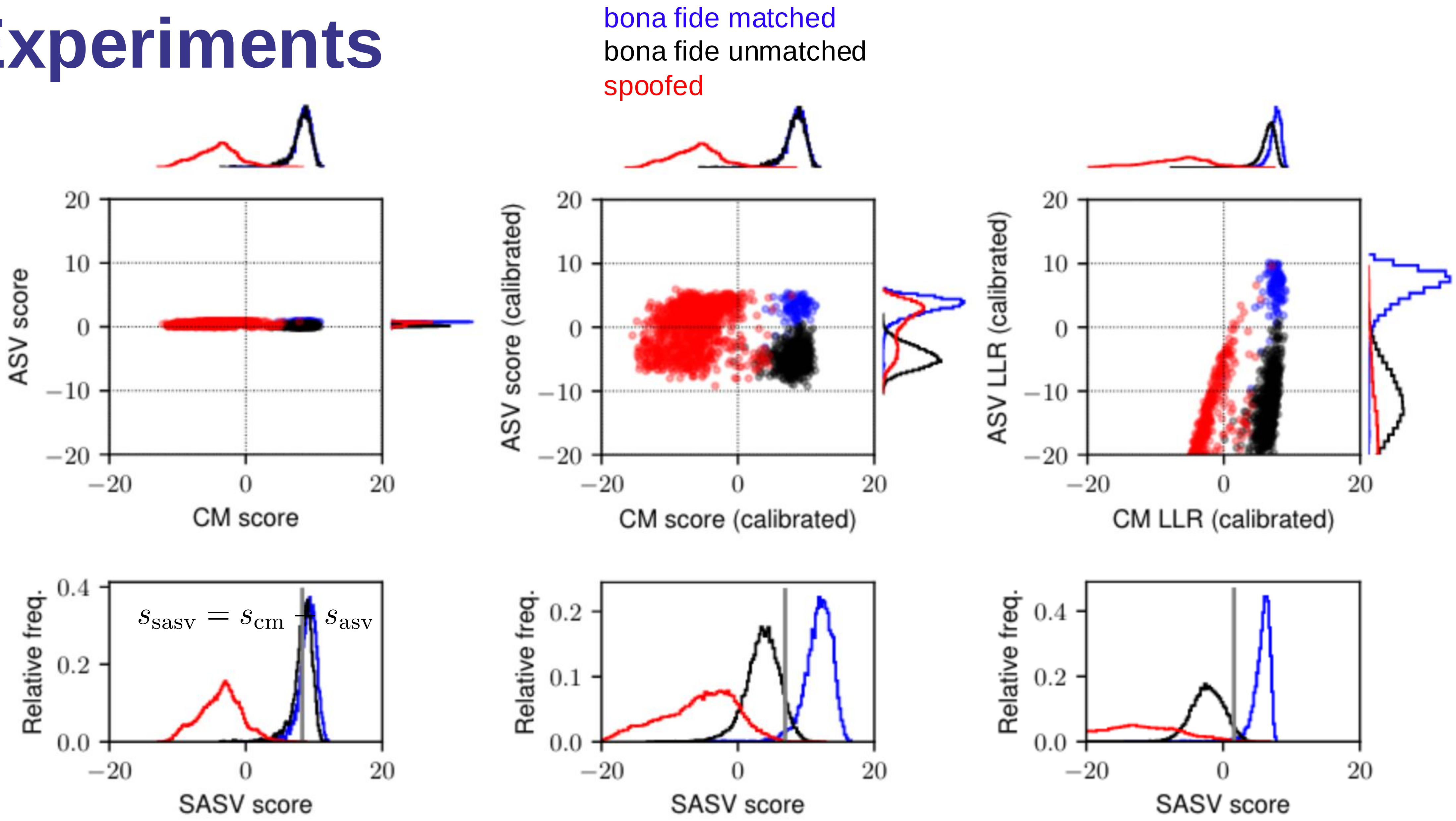
baseline good linear fusion good linear fusion

Experiments



baseline | good linear fusion | good linear fusion

Experiments



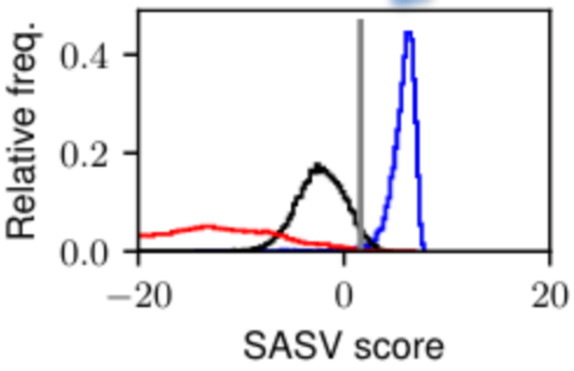
NII

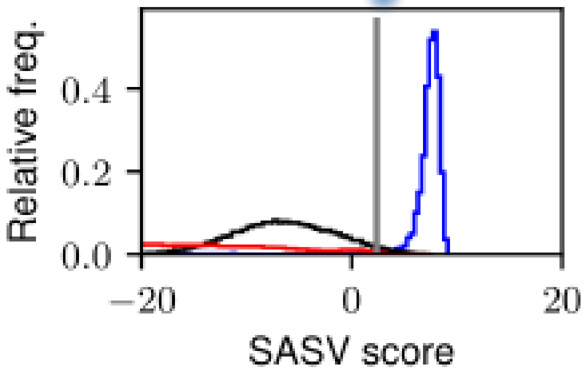
Experiments

ID	B1	B1c	L2	L2c	L3	L3c	B1v2	Post
Fusion				linear		non-liear	(Jung 2022)	(Zhang 2022)
Calibration	×	✓	×	✓	×	✓	×	×
SASV-EER (%)	20.46	2.73	3.31	1.56	1.44	1.43	1.60	1.55
conf. ($\alpha = 5\%$)	±0.40	±0.27	±0.31	±0.23	±0.23	±0.23	±0.22	±0.24

The difference is small on this database

good linear fusion



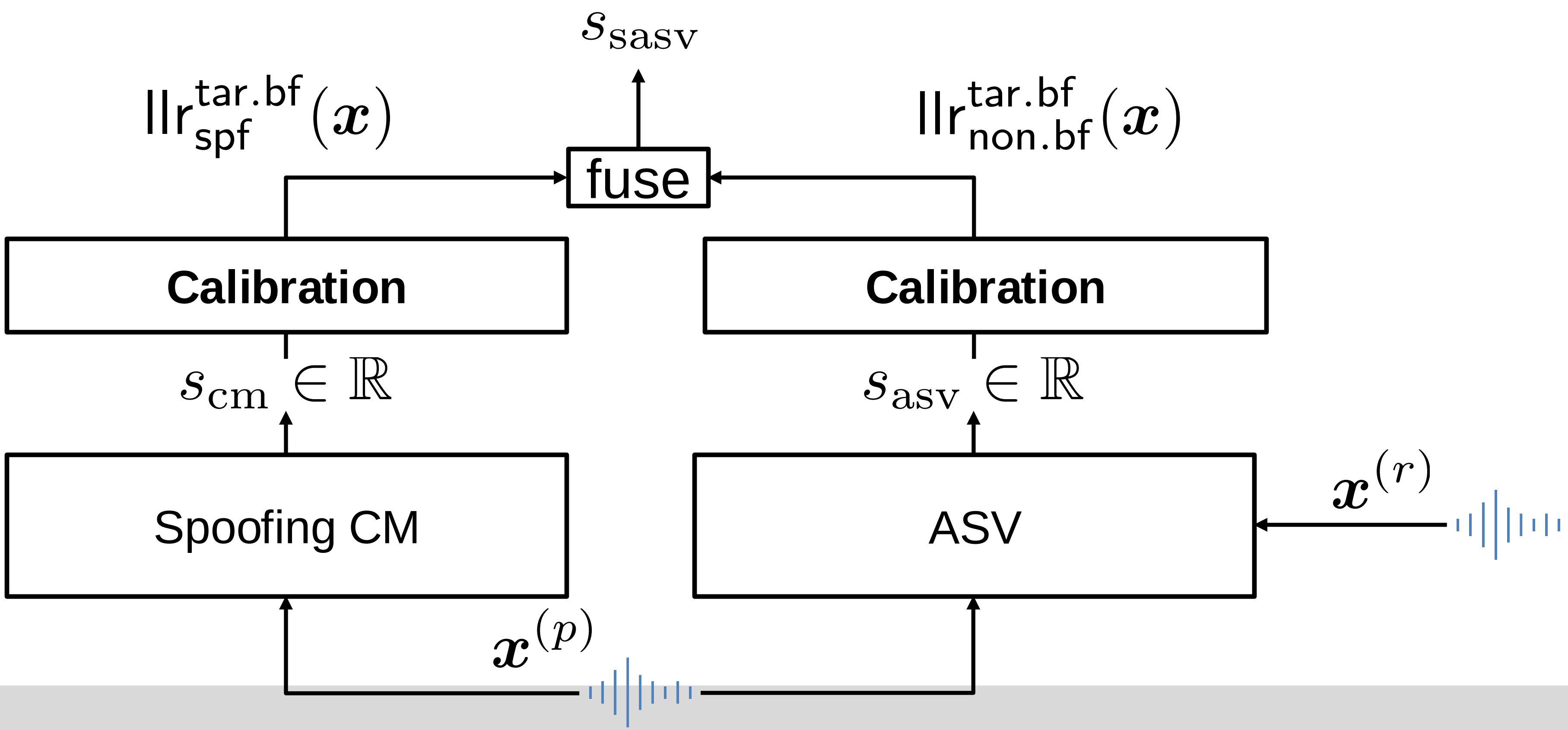


good non-linear fusion

Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. 2022. SASV 2022: The first spoofing-aware speaker verification challenge. In Proc. Interspeech, 2022. 2893–2897.
You Zhang, Ge Zhu, and Zhiyao Duan. 2022. A Probabilistic Fusion Framework for Spoofing Aware Speaker Verification. In *Proc. Odyssey*, June 28, 2022. ISCA, 77–84.

Main messages

- ❑ Fusion SASV != fusion of ASV or CM ensemble
- ❑ Linear and non-linear can be suppored by theory
- ❑ Calibration affects discrimination



Pointers

- **Evaluation using the same Bayes decision cost**

  Hye-jin Shim, Jee-weon Jung, Tomi Kinnunen, Nicholas Evans, Jean-Francois Bonastre, and Itshak Lapidot. 2024. **a-DCF: an architecture agnostic metric with application to spoofing-robust speaker verification**. In Proc. Odyssey, 2024. 158–164. https://doi.org/10.21437/odyssey.2024-23

- **SOTA ASV is not robust to spoofing attacks**

  Jee-weon Jung, Xin Wang, Nicholas Evans, Shinji Watanabe, Hye-jin Shim, Hemlata Tak, Sidhhant Arora, Junichi Yamagishi, and Joon Son Chung. 2024. **To what extent can ASV systems naturally defend against spoofing attacks?** In Proc. Interspeech, 2024. .

  A4-05.5

- **The non-linear fusion has been used by many teams in ASVspoof 5 challenge**

Thank you

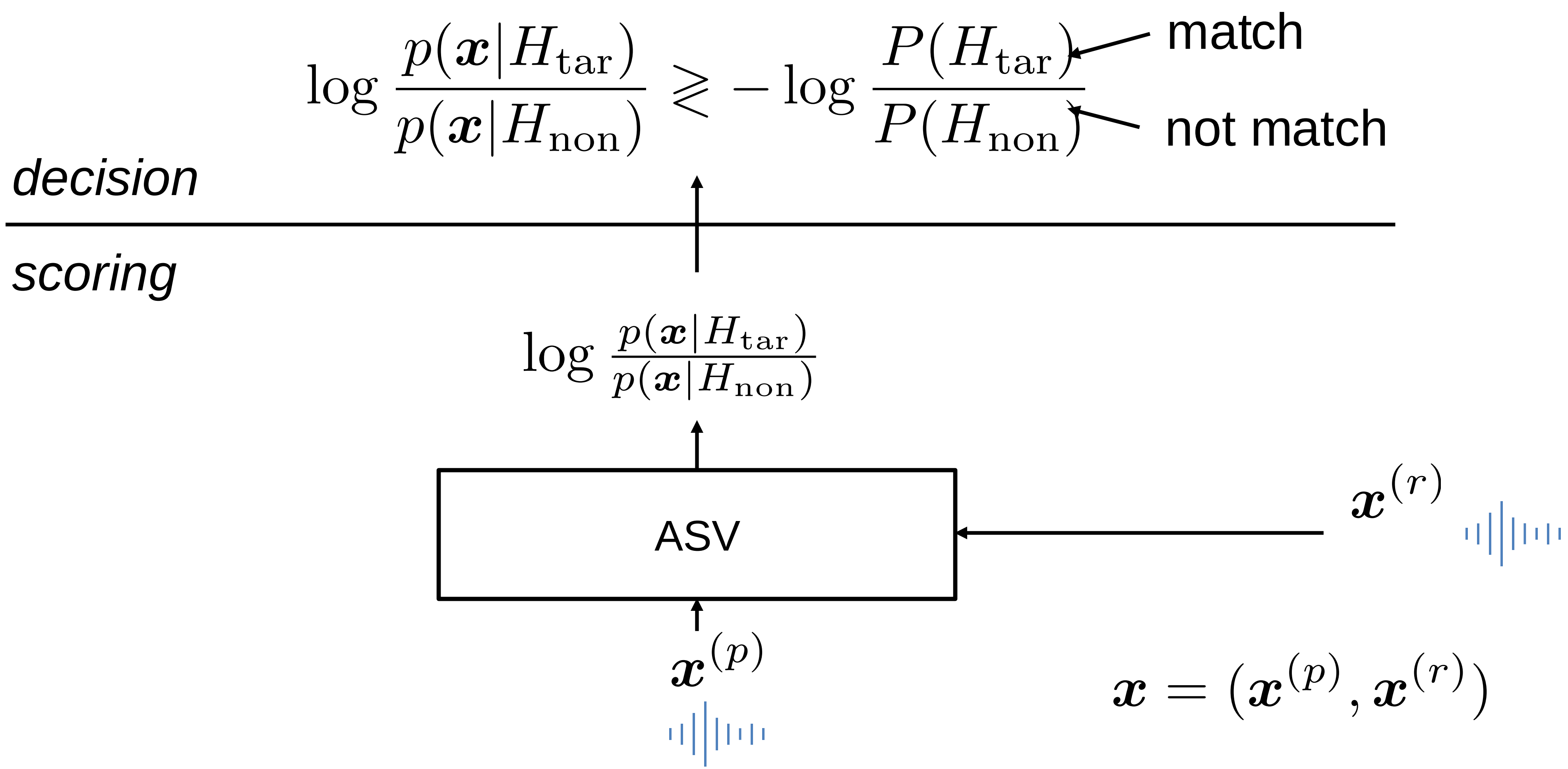
Code & Jupyter notebook
step-by-step explanation

Appendix
theory in details

ASVspoof

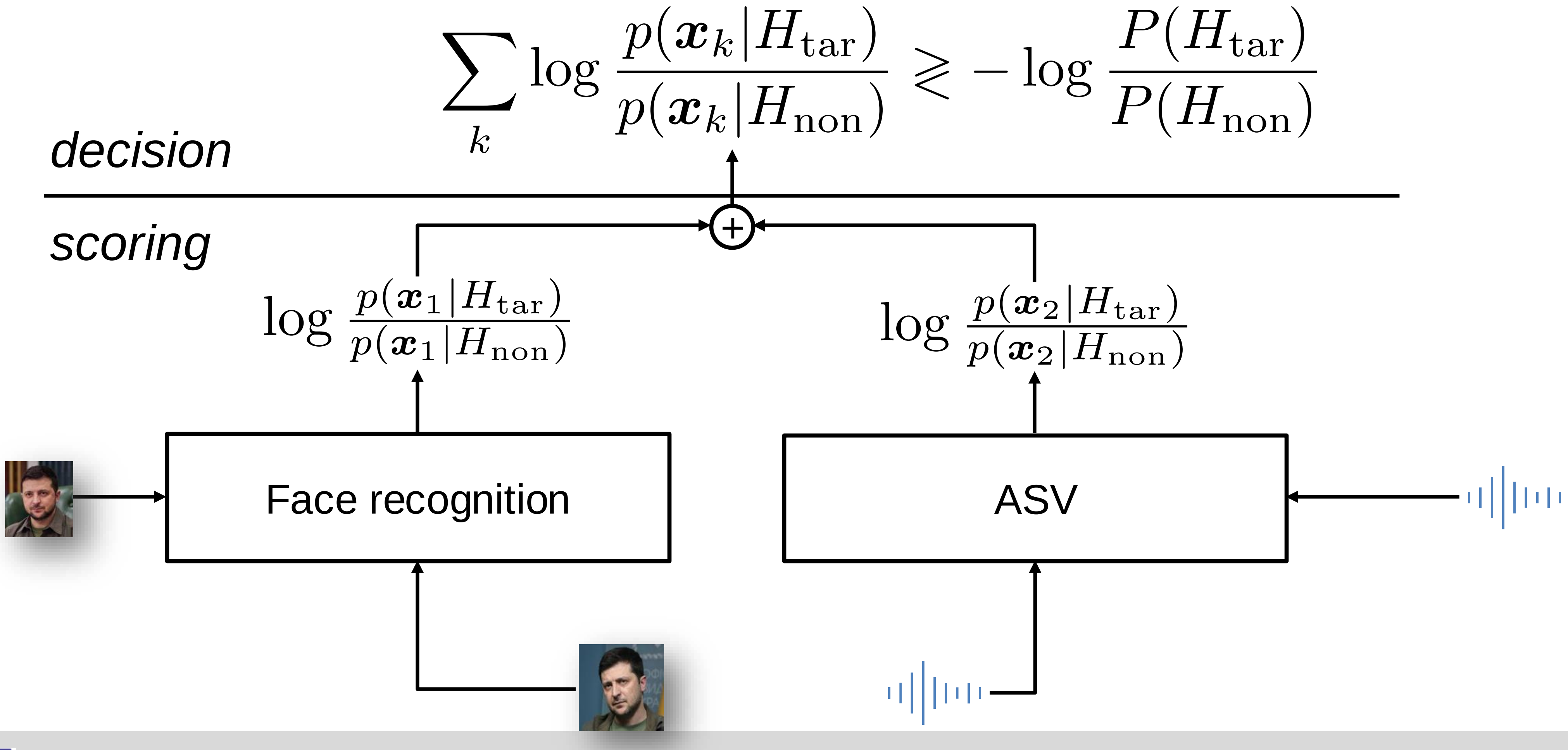
Fusing CM & ASV is special

❑ A single ASV



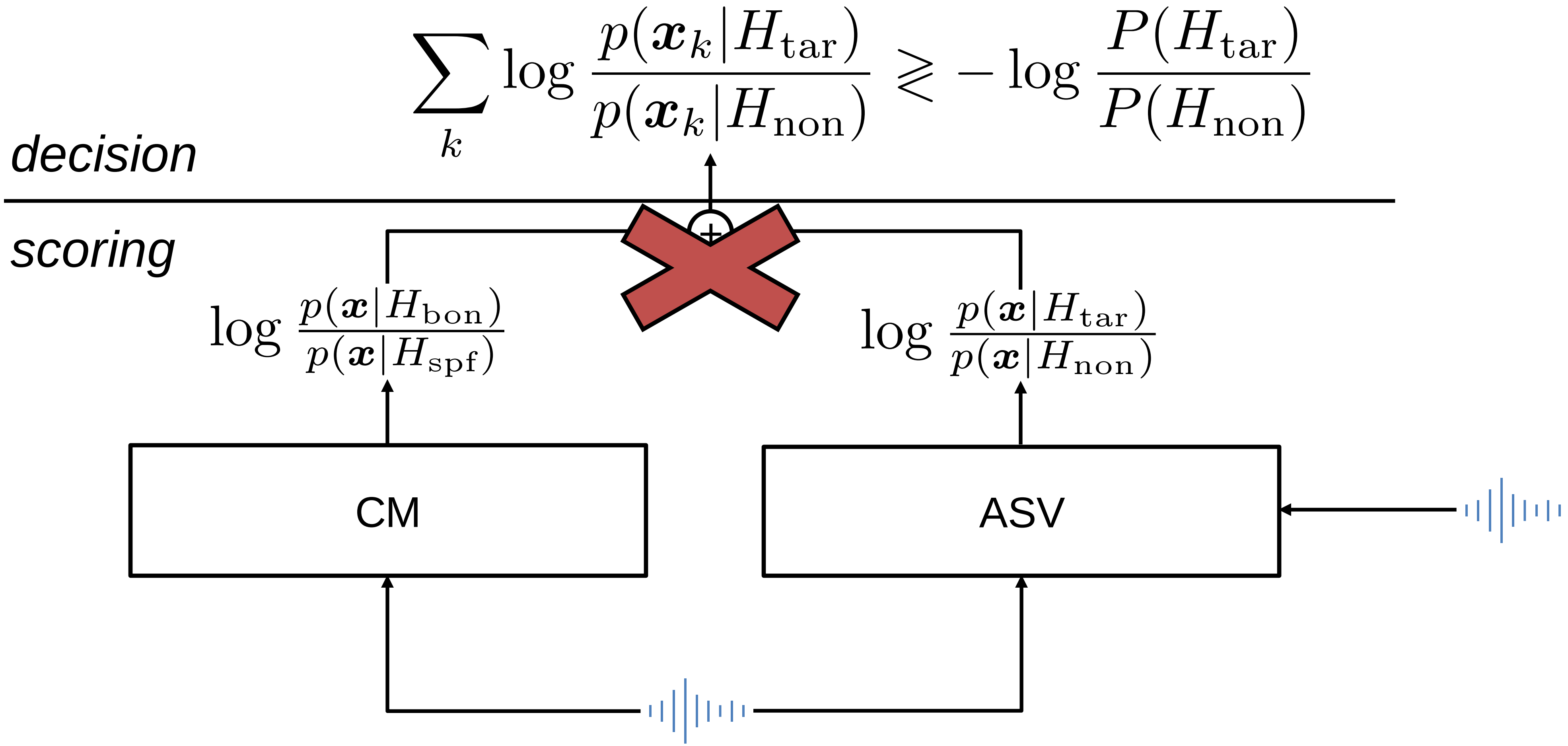
Fusing CM & ASV is special

- Fusing ASV, face recognition, and other biometrics



Fusing CM & ASV is special

- CM and ASV are dealing with different hypotheses

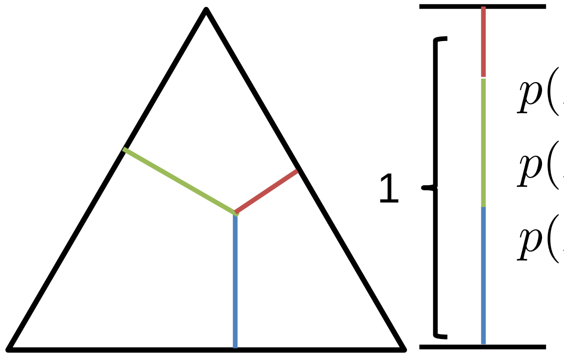


Fusing CM & ASV is special

- ❑ **We have three classes of data in two separate hypothesis testings**

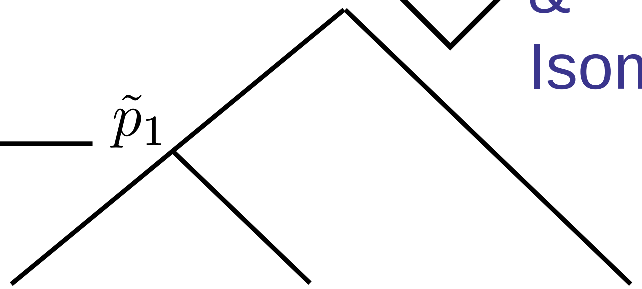
$$\{H_{\text{fake}}, H_{\text{real-diff}}, H_{\text{real-match}}\}$$

Simplex



Bayes' rule
&
Isometric-log-ratio

What we need



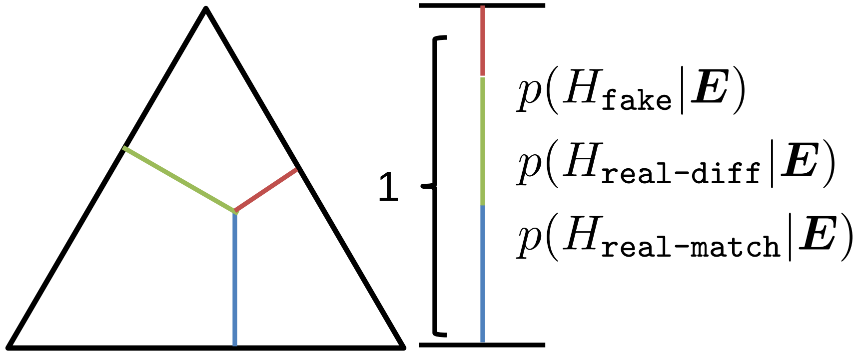
Optimal way using ternary hypothesis testing

Fusing CM & ASV is special

- We have three classes of data in two separate hypothesis testings

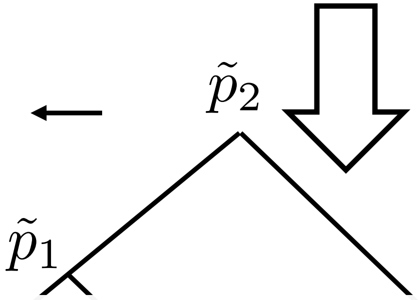
$$\{H_{\text{fake}}, H_{\text{real-diff}}, H_{\text{real-match}}\}$$

Simplex



Bayes' rule
&
Isometric-log-ratio

$$\tilde{p}_2 = \frac{1}{\sqrt{6}}\left[\log\frac{p(\boldsymbol{E}|H_{\texttt{real-match}})}{p(\boldsymbol{E}|H_{\texttt{fake}})} + \log\frac{p(\boldsymbol{E}|H_{\texttt{real-match}})}{p(\boldsymbol{E}|H_{\texttt{real-diff}})}\right] \leftarrow$$



log likelihood ratio   log likelihood ratio