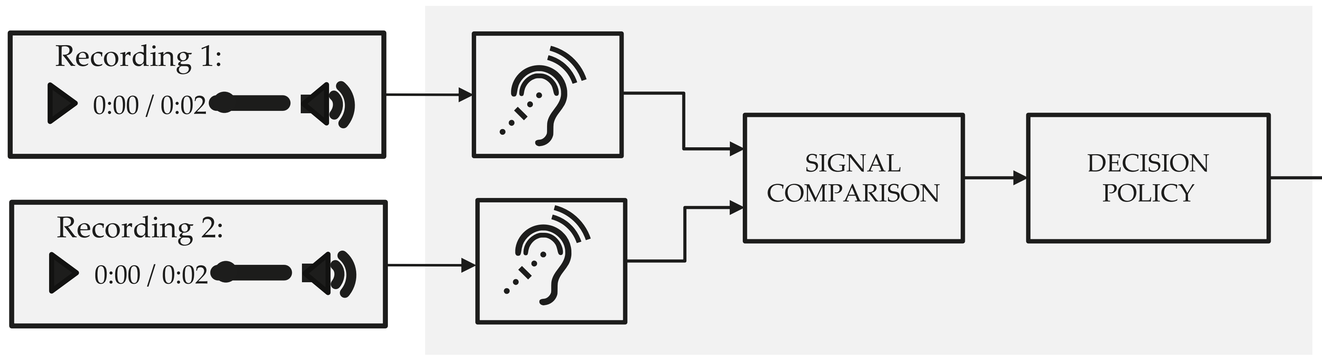
Speaker Detection by the Individual Listener and the Crowd: Parametric Models Applicable to Bonafide and Deepfake Speech

Tomi Kinnunen[1], Rosa González Hautamäki[2,1]
Xin Wang[3], Junichi Yamagishi[3]

[1]University of Eastern Finland, [2]University of Oulu, [3]National Institute of Informatics

Speaker recognition by humans is relevant to (1) neurosciences, (2) human-in-the-loop applications, (3) evaluation of speech synthesis. Often data is analyzed using nonparametric methods. Our work seeks to remind about the half-forgotten **signal detection theory** for parameteric modeling of speaker detection by (groups of) listeners, demonstrated through two case studies.

Modeling speaker detection ("same or different speaker?") by listeners



i: listener id, *t*: trial id

$$Y_{i,t}^{\text{tar}} = \begin{cases} 1 & \text{if } S_{i,t} + b_i \geq 0 \\ 0 & \text{if } S_{i,t} + b_i < 0 \end{cases}$$

$S_{i,t}$: speaker similarity score, b_i: bias

A generalized linear mixed effects (GLME) model

"The Crowd" part: generative between-listener model of the detection model parameters

$$\begin{bmatrix} b_i \\ d_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} b + b^{(k_i)} \\ d + d^{(k_i)} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix} \right)$$

k_i : group id
σ_u^2 σ_v^2 : between-listener variances of *b* and *d*
ρ : correlation of *b* and *d*

"The individual listener" part: parametric detection model for listener *i*

$$S_{i,t} = d_i \, X_t^{\text{tar}} + \varepsilon_{i,t}$$

d_i : discrimination parameter
X_t^{tar} : ground-truth (0 or 1)
$\varepsilon_{i,t}$: residual (here, either normal or logistic)

➔ Miss and false alarm rates (under logistic model)

$$P_{\text{miss}}^{(i)} = 1 - \Pr\left(Y_{i,t} = 1 | X_t^{\text{tar}} = 1\right) = \left(1 + e^{b_i + d_i}\right)^{-1}$$

$$P_{\text{fa}}^{(i)} = \Pr\left(Y_{i,t} = 1 | X_t^{\text{tar}} = 0\right) = \left(1 + e^{-b_i}\right)^{-1},$$

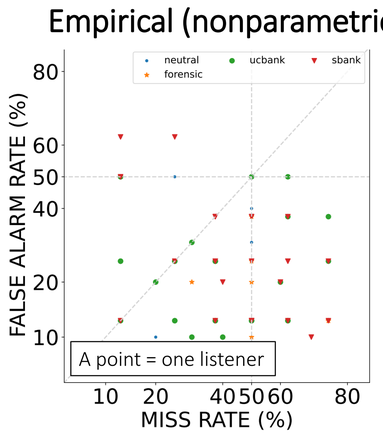
Case study I: impact of role-play

100 target + 100 nontarget trials from VoxCeleb1-E, AMT experiment: 200 trials x 5 repetitions of each x 4 groups = 4000 ratings, 60 listeners (disjoint sets) per panel (4 x 60 = 240 listeners)

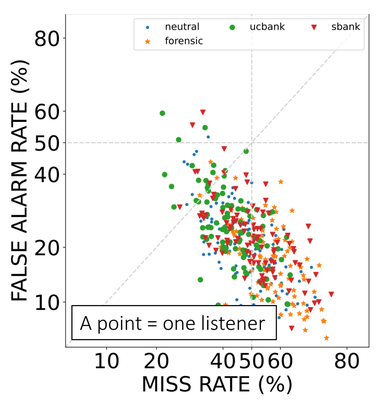
	Probit		Logit	
Scenario	$b+b^{(k_i)}$	$d+d^{(k_i)}$	$b+b^{(k_i)}$	$d+d^{(k_i)}$
Neutral	-1.352	1.484	-2.349	2.558
Forensic	-1.563	1.258	-2.757	2.274
User-c. bank	-1.149	1.407	-1.970	2.390
Secure bank	-1.321	1.230	-2.272	2.126

Corr. of b and d: $\rho = -0.66$ $\rho = -0.75$

Empirical (nonparametric)



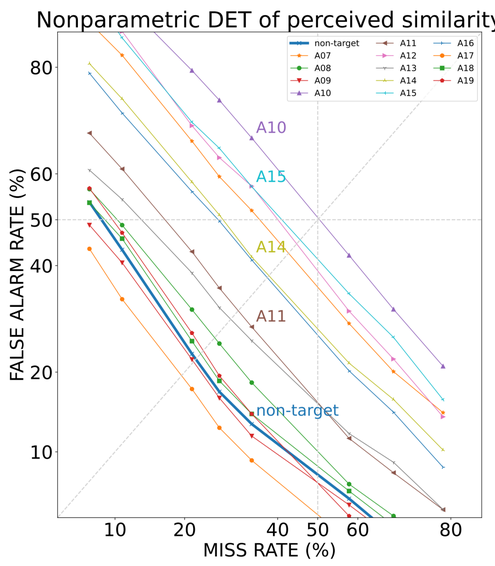
Simulated from the model



Case study II: potential to simplify similarity rating scale

Speaker similarity ratings in ASVspoof 2019 data (deepfake / spoofed speaker identities). 13 different spoofing methods, rated by 1145 listeners using a 10-point scales → converted to binary decisions for our model (1…5→0; 6..10→1)

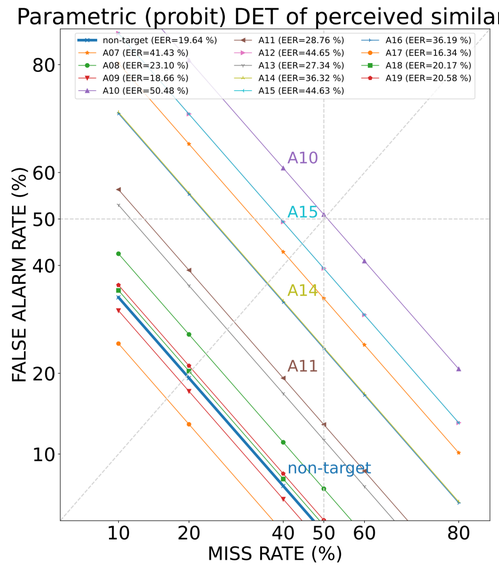
Nonparametric DET of perceived similarity



Nonparametric DETs based on 10-point rating scale

≈

Parametric (probit) DET of perceived similarity



Parametric DETs fitted to binarized listener responses.