La voix, une modalité d'identification unique

avec ses limites et ses risques

Journées Nationales 2024 du GDR Sécurité Informatique

Jean-François Bonastre

Inria DEES & LIA Avignon Université

Jean-francois.bonastre@inria.fr

Inria

Défense&Sécurité

I - Introduction

La voix, une modalité d'identification unique

- Présence fréquente
  - Souvent le seul média disponible
- Facilité de captation
- Utilisation « mains libres »
- Acceptabilité
- Aspect « biométrique »
- Performances alléchantes

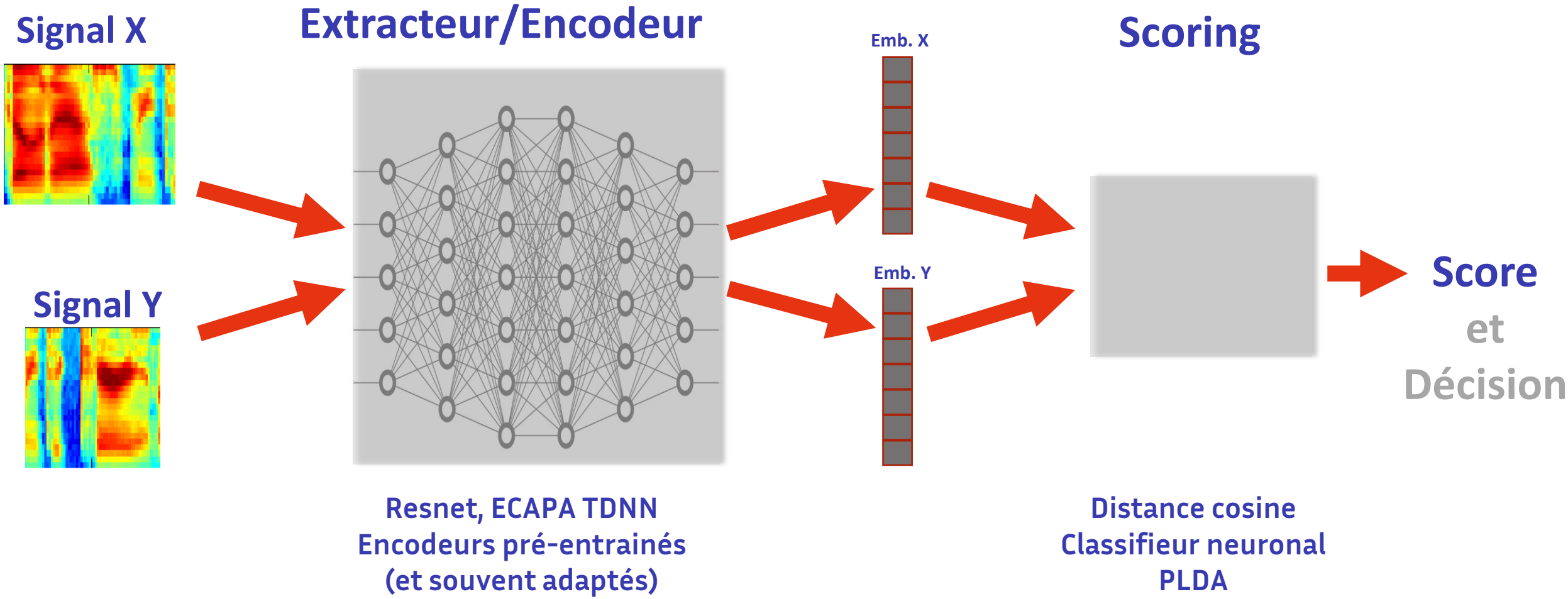
- Facilité d'utilisation & convivialité
  - Mode de communication naturel, sans apprentissage spécifique
  - Peut s'employer sans protocole spécifique, durant une conversation ou une interaction personne-machine (ou sans consentement…)
  - Peut être associer à un protocole « à secret » (convivial)

Reconnaissance du locuteur (RAL)

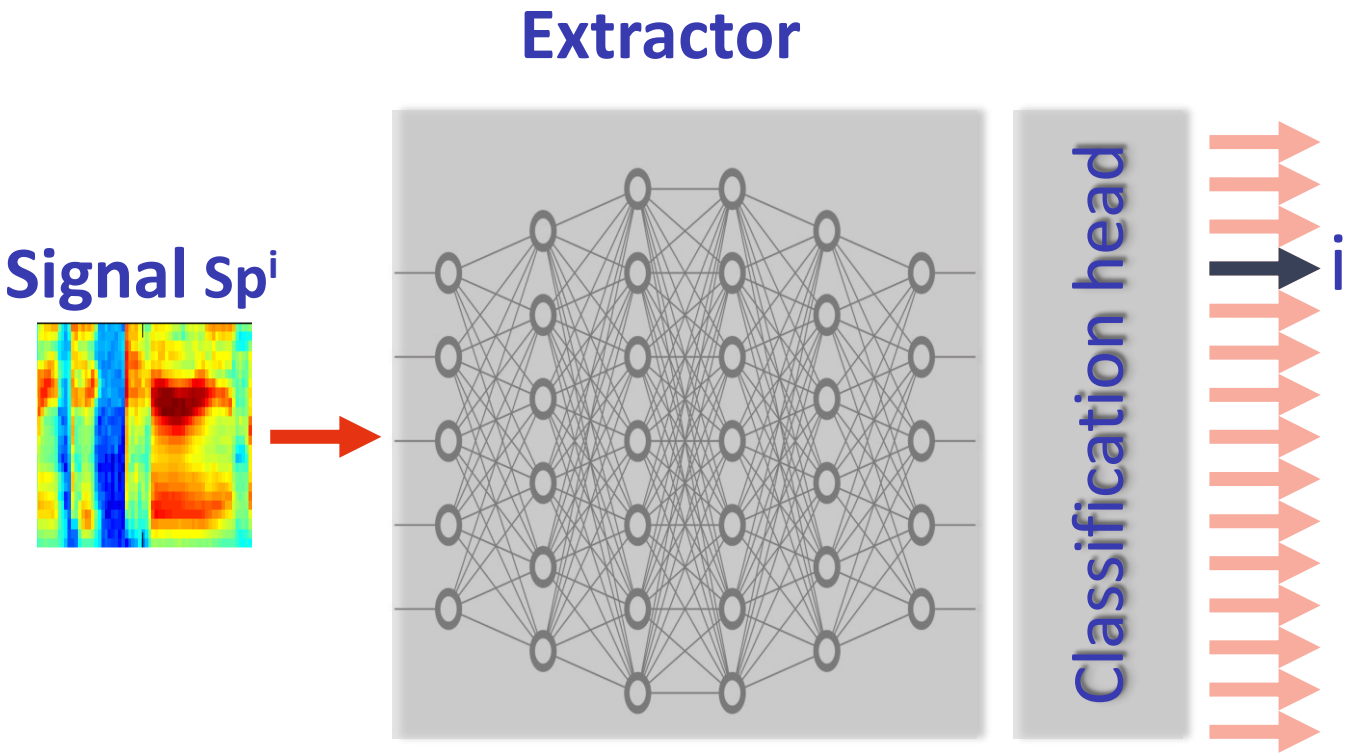
- La RAL consiste à décider si un extrait de parole, Y, a bien été prononcé par un locuteur donné, X
- X peut être représenté par
  - Un modèle de sa voix appris à partir de plusieurs enregistrements (avec contrôle)
  - Ou un enregistrement de sa parole = « comparaison de voix »

?

Reconnaissance du locuteur (RAL)



Reconnaissance du locuteur (RAL)



- Tâche de classification fermée
  - ~6000 locuteurs
  - ~1 million d'enregistrements
  - « Data augmentation »
  - Le modèle de l'extracteur possède ~100 millions de paramètres
- Le « scoring » est souvent basé sur une simple fonction cosinus (est-ce une conséquence ?)

Fonctionne plutôt très bien (~0.9 % EER sur VoxCeleb)

Fonctionne (très) bien mais…

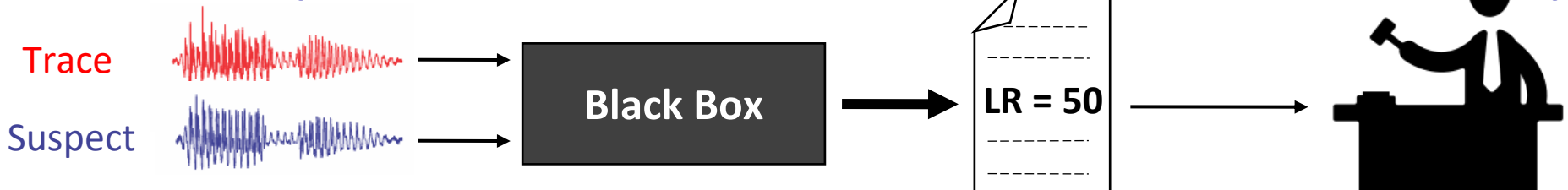
- **En termes de performance, cela reste une modalité « moyenne » et avec une GRANDE variabilité dans les résultats** :
  - En fct de l'extrait de référence (*J.Kahn and al, 2010*): **7** x **EER**
  - Ou du contenu phonétique (M. Ajili and al., 2016)
  - Changement de condition (*C.S Greenberg and al, 2011*), **5** x **EER**
  - Doddington '' zoo": **sheep, goats, lambs, wolves**
  - Une grande variabilité : style de parole, bruits, environnements inconnus, langues peu présentes, dialectes, accents…
- **Des applications souvent sensibles (justice, défense&sécurité)**
- **La voix contient beaucoup plus que l'identité**
- **Les attaques, notamment par « deepfake », sont à prendre en considération**

Fonctionne (très) bien mais…

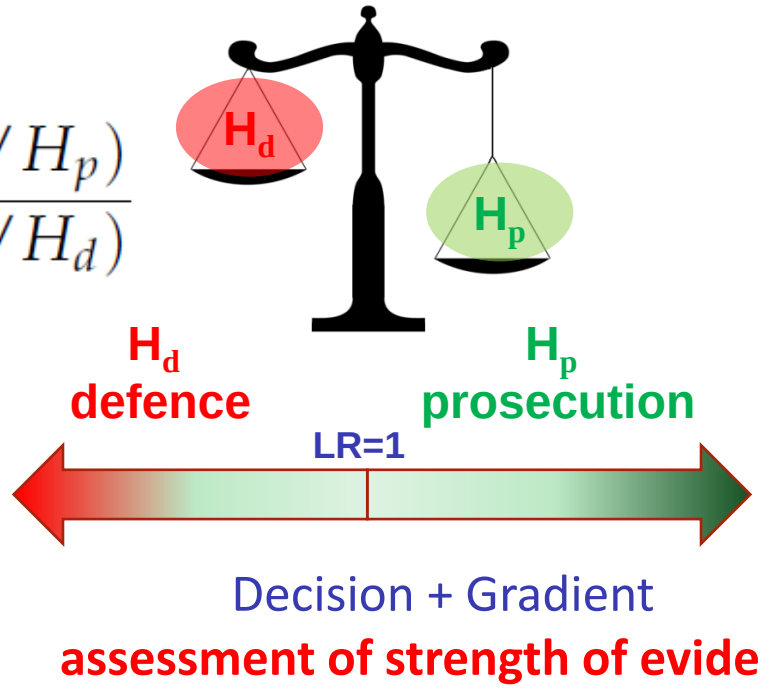
- Besoin d'explicabilité
  - Confiance, fiabilité et interprétation
  - Reconnaitre/traiter les biais
  - Evolution réglementaire
- Besoin de se défendre face aux attaques
  - Ajout de contremesures
  - (Explicabilité)
- Besoin de protéger la vie privée
  - Explicabilité (savoir ce qui est dévoilé/utilisé)
  - Anonymisation différentielle de la voix, pour ne dévoiler que ce qui est utile

II - Besoin d'explicabilité Modélisation par attributs partagés (BA-LR)

Un cas d'école :
La comparaison de voix en criminalistique



$$LR = \frac{p(E/H_p)}{p(E/H_d)}$$



- Incertitude ?
- Basé sur quoi ?
- Quels facteurs amènent ce LR ?
- Quelles caractéristiques (fact.) ?

➔ **Avec seulement le LR, il y a un manque clair de transparence !**

Un cas d'école :
La comparaison de voix en criminalistique

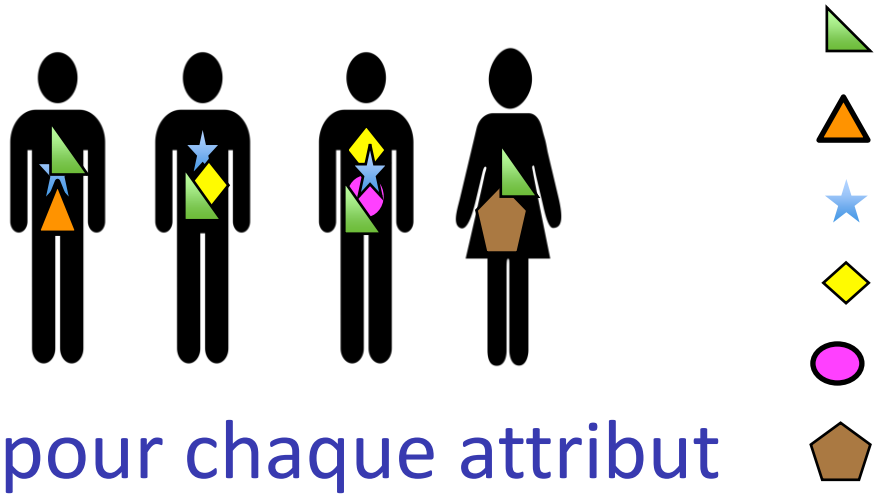
- Une situation pas toujours simple en France…
- Moratoire
- Intervention des scientifiques dans les tribunaux comme « témoins »
- Travaux scientifiques avec les laboratoires de police scientifique nationaux
- Bonastre, J. F. (2020). 1990-2020: retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)* (pp. 38-47). ATALA; AFCP.

Binary-Attribute-based Likelihood Ratio (BA-LR)

Un autre paradigme de représentation

1. Modéliser un locuteur par un ensemble d'attributs partagés par un groupe de locuteurs

BA-LR



2. Calculer un LR partiel pour chaque attribut
2. Basé sur des caractéristiques de l'attribut compréhensibles par tous
3. « Interpréter » les attributs

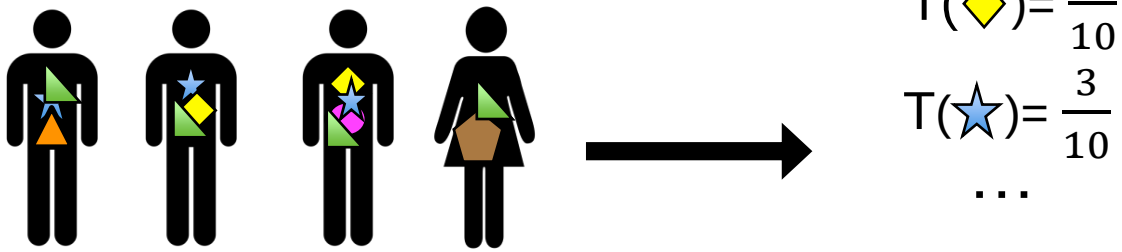
Binary-Attribute-based LR (BA-LR)
1: Binary Attribute (BA) Extraction

- Une première implémentation
  - Adaptation « légère » du baseline resnet (256 dim) par petites modifications pour favoriser la binarisation (softplus / no batch norm) & Apprentissage « classique »
  - Binarisation ultrasimple : si x~=0, f(x)=0, sinon f(x)=1
  - Pas de garantie d'indépendance entre les attributs
  - Suppression des BAs inactives -> vecteur BA de 205 dimensions
- Deuxième version : encodeur classique suivi d'un auto-encodeur binaire
  - Binarisation intrinsèque
  - Ajout d'une fonction de coût spécifique à la notion d'attribut (logique contrastive learning)
  - Taille du vecteur choisie indépendamment de l'encodeur
  - Fine tuning de l'encodeur : à venir

Binary-Attribute-based LR (BA-LR)
2: BA paramètres comportementaux explicites

Paramètres comportementaux appris sur une population de référence

Typicalité du BA_i (T_i): fréquence de couples partageant l'attribut dans la population



$T(\diamond) = \frac{1}{10}$

$T(\star) = \frac{3}{10}$

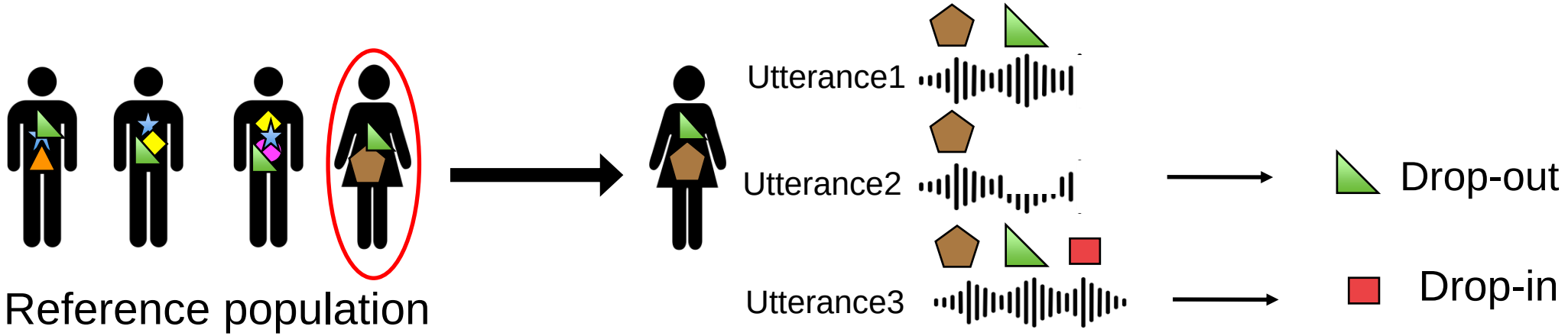
...

$T(\blacktriangleright) = \frac{6}{10}$

Binary-Attribute-based LR (BA-LR)
2: BA paramètres comportementaux explicites

Dropout $(Dout_i)$: Probabilité pour qu'un attribut ne soit pas détecté dans un enregistrement d'un locuteur qui a montré cet attribute au moins une fois

Dropin (Din) : Probabilité de détecter à tord un attribute, à cause d'un bruit par exemple (incertitude) . Fixé par heuristique



Binary-Attribute-based LR (BA-LR) 2: LR estimation (scoring explicite)

- Deux versions basées sur le même principe :
  - LR partiel par attribut estimé par écriture directe des hypothèses accusation et défense en probabilités estimées par la présence et les paramètres des BAs
  - LR final est obtenu comme le produit des LRs partiels en assumant l'indépendance entre les BAs
- Version 1, directement inspiré de la ref., avec la notion de « profil » & « trace »
- Version 2, comparaison de deux « traces »

Inspired from:
P. Gill et al., "Interpretation of complex dna profiles using tippett plots," Forensic science international: Genetics supplement series 1, pp. 646–648, 2007.
P. Gill et al., "Dna commission of the international society of forensic genetics: Recommendations on the evaluation of str typing results that may include drop-out and/or drop-in using probabilistic methods," Forensic Sci Int Genet, pp. 679–688, 2012.
T. Tvedebrink et al., "Estimating the probability of allelic drop-out of str alleles in forensic genetics," Forensic Sci Int Genet, pp. 679–688, 2012.
D. J. Balding and J. Buckleton, "Interpreting low template dna profile," National library of medicine, 2010.

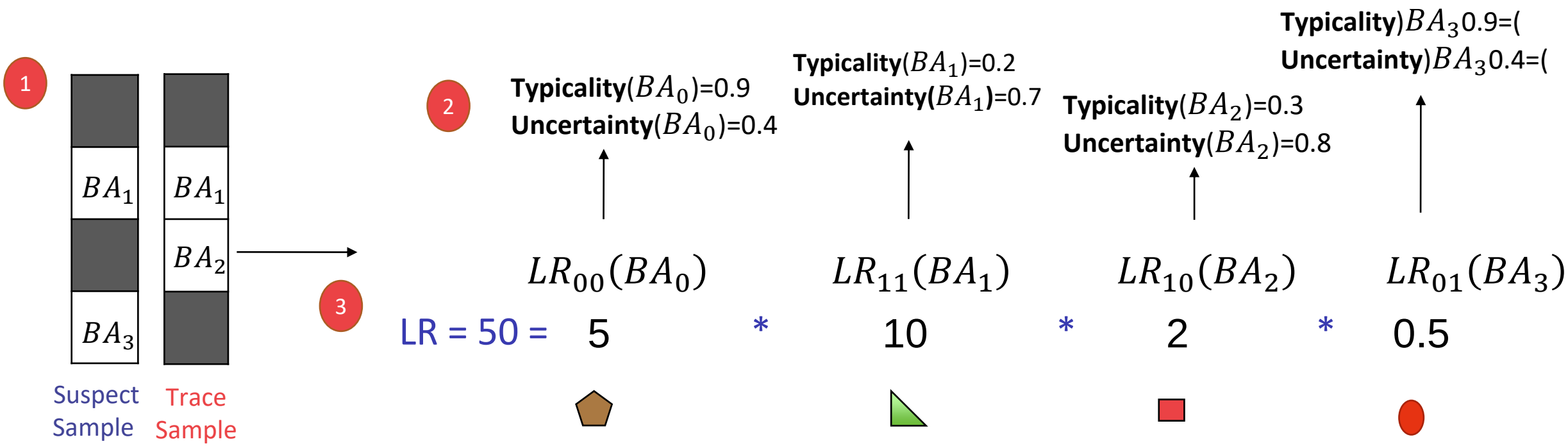
Binary-Attribute-based LR (BA-LR) 2: LR estimation (scoring explicite)

- Pour X, Y et un attribut I, le LR partiel est basé uniquement sur
  - BA^x_i et BA^y_i (00, 10, 11, 01)
  - BA_i paramètres comportementaux
  - (LR_{10} et LR_{01} sont symétriques par moyennage)

$$lr(\mathrm{BA}_i) = \begin{cases} \dfrac{\mathrm{Dout}_i}{T_i \cdot (\mathrm{Dout}_i + \overline{\mathrm{Dout}_i})} & \text{if } \mathrm{Y}(\mathrm{BA}_i = 0), \mathrm{X}(\mathrm{BA}_i = 1) \\ \dfrac{1}{T_i \cdot (\overline{\mathrm{Din}} + \mathrm{Dout}_i)} & \text{if } \mathrm{Y}(\mathrm{BA}_i = 0), \mathrm{X}(\mathrm{BA}_i = 0) \\ \dfrac{1}{T_i \cdot (\overline{\mathrm{Dout}_i} + \mathrm{Din} \cdot T_i)} & \text{if } \mathrm{Y}(\mathrm{BA}_i = 1), \mathrm{X}(\mathrm{BA}_i = 1) \\ \dfrac{\mathrm{Din} \cdot T_i}{T_i \cdot (\mathrm{Din} \cdot T_i + \overline{\mathrm{Din}})} & \text{if } \mathrm{Y}(\mathrm{BA}_i = 1), \mathrm{X}(\mathrm{BA}_i = 0) \end{cases}$$

- $\overline{Din} = 1 - Din$
- $\overline{Dout} = 1 - Dout$

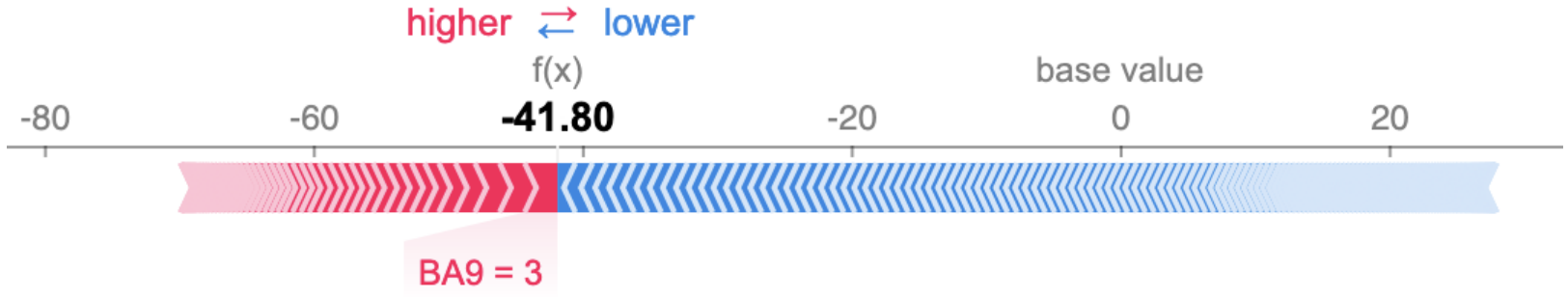
Binary-Attribute-based LR (BA-LR)
2: LR estimation (scoring explicite)



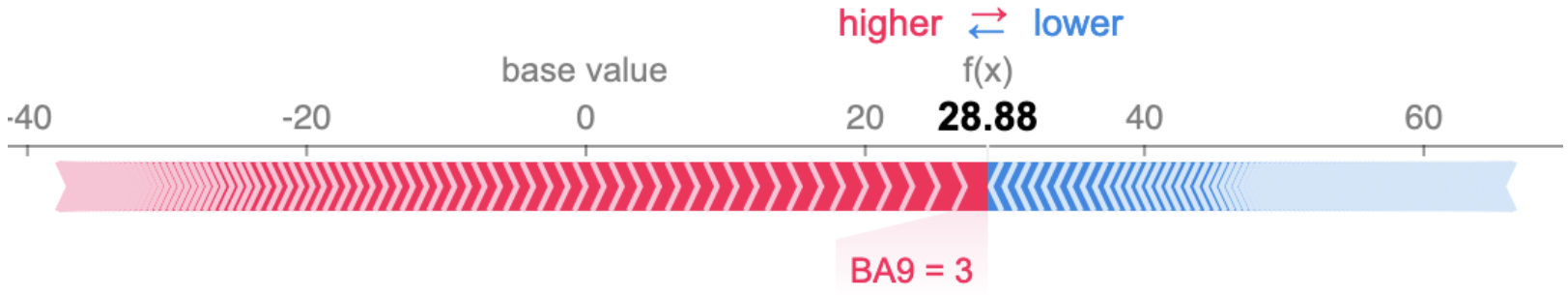
BA-LR: Interprétabilité locale

(using SHAP toolkit)

Non target pair



Target pair



BA	BA207	BA9	BA255
Contribution	1.93	2.44	1.86
Typicality	0.565	0.156	0.582
Dropout	0.783	0.458	0.775
Cat pLLR	11	11	11

Binary-Attribute-based LR (BA-LR)
3: Interprétabilité des attributs

- Logique « informative world » et « mapping function »
- Espace de représentation « informatif » réalisé en extrayant des paramètres phonétiques/acoustiques sur chaque exemple (F0, Formants, Cepstres, RSB, …). Choix d'une partie d'opensmile pouvant être étendu
- Pour chaque attribut
  - répartition des exemples d'entrainement en deux ensembles, 1(présence), 0(exemples provenant de locuteurs n'ayant JAMAIS montré l'attribut)
  - Explication de la différence entre les 2 ensembles en utlisant l'espace informatif par decisionTree/Shap (ou StepwiseLDA)
  - Sélection des variables significatives

- (résultats dans les dernières publications)
- Extension à l'analyse segmentale (localisation des informations)
- Extension (nvle thèse) : interprétabilité par des experts en phonétique / perception

Conclusion : BA-LR, une voie pour une approche intrinsèquement explicable

- Jeu d'attributs préétablis et explicables
- Scoring explicable (optionnel)
- Performance acceptable (~2% de perte en EER) et en progrès
- Généralisation testée avec succès sur d'autres corpus/langues (autres tâches comme la détection de la langue en cours)

III – Attaques et contremesures

Une problématique forte

- Déjà ancienne (Bonastre, J. F., Matrouf, D., & Fredouille, C. (2007). Artificial impostor voice transformation effects on false acceptance rates. In *Interspeech*.)
- Mais qui prend une très grande importance dans le cadre de la lute anti-influence
- L'exemple de ASVspoof
  - Suite de challenges (Interspeech) depuis 2017. Eurecom en France, UEF (Finlande) et NII (Japon)
  - ASVspoof 2021
    - Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., ... & Delgado, H. (2021, September). ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*.
    - De TRES bon résultats
    - Mais des questions sur la réalité des scénarios d'attaque
  - ASVspoof 5 (2024) en cours (https://www.asvspoof.org/)
    - Un scénario (un peu) plus poussé
- Défi DGA 2024

IV – Protection de la vie privée et Anonymisation différentielle de la voix

L’exemple du challenge VoicePrivacy

- Initiative LIA, Eurecom, Multispeech Inria, NII
  - Mise en commun d’efforts de divers projets dont ANR-JST Franco-Japonais VoicePersonae (LIA, Eurecom, NII)
- Quelles métriques ?
  - Noé, P. G., Bonastre, J. F., Matrouf, D., Tomashenko, N., Nautsch, A., & Evans, N. (2020, October). Speech Pseudonymisation Assessment Using Voice Similarity Matrices. In *Interspeech 2020*.
  - Nautsch, A., Patino, J., Tomashenko, N., Yamagishi, J., Noé, P. G., Bonastre, J. F., ... & Evans, N. (2021, October). The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment. In *Interspeech 2020* (pp. 1698-1702). ISCA.
- Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P. G., ... & Maouche, M. (2022). The voiceprivacy 2020 challenge: Results and findings. *Computer Speech & Language*, *74*, 101362.

Une extension, la notion de « perfect privacy »

- Thèse de Paul-Gauthier Noé
- Extension de « perfect secret »
- Vue bayésienne : dévoiler les données ne doit pas changer le rapport de vraisemblance (LR)
- Anonymisation dans l'espace des LLR
- Retour dans l'espace des embeddings par un mapping bidirectionnel « Normalizing Flow »
  - Noé, P. G., Nautsch, A., Matrouf, D., Bousquet, P. M., & Bonastre, J. F. (2022, May). A bridge between features and evidence for binary attribute-driven perfect privacy. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3094-3098). IEEE.
- Extension à la perception par resynthèse
  - Noé, P. G., Miao, X., Wang, X., Yamagishi, J., Bonastre, J. F., & Matrouf, D. (2023, June). Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
- Extension au cas multiclasse et proposition de l'extension multiclasse du LLR
  - Noé, P. G. (2023). *Representing evidence for attribute privacy: bayesian updating, compositional evidence and calibration* (Doctoral dissertation, Université d'Avignon).

Conclusion

- En termes d'authentification des utilisateurs, la voix
  - Est un media d 'intérêt, ne serait-ce que parce que souvent elle est le seul média disponible
  - Pour un niveau de performance/sécurité élevé, la « voix » sera rarement utilisée seule (ex. + « secret » obtenu par dialogue)
- La voix peut être liée à des applications sensibles, dans le cadre judiciaire ou défense/sécurité
- La voix transporte des informations personnelles
  - Qui peuvent être exploitées utilement (détection enfant/adulte, stress/fatigue)
  - Qui impliquent cependant un risque de biais, de discrimination, pouvant nécessiter d'être protégées pour des raisons de « privacy »
- Avenir ?
  - Explicabilité/Interprétabilité et fiabilité + que performance
  - Anonymisation différentiée
  - Détection des attaques/deepfakes (Une bataille encore gagnable ?)

ATTENTION A LA GENERALISATION :

- La détection en milieu ouvert est difficile à évaluer
- Un besoin de précaution est indispensable au vu des conséquences « humaines » (Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J. F., & Matrouf, D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, *26*(2), 95-103.)

Merci & ref

- **Chaire LIAvignon (PhD sponsor), liavignon.fr**
- **Imen Ben-Amor**
  - Thèse « Modélisation profonde basée sur la notion d'attributs de voix pour la reconnaissance du locuteur explicable : application au domaine criminalistique », 25/04/2024
  - Speaker Odyssey 2024 & Interspeech 2024
  - Ben-Amor, I., Bonastre, J. F., O'Brien, B., & Bousquet, P. M. (2023, August). Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition. In *Interspeech 2023*.
  - Imen Ben Amor, Jean-François Bonastre: BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison, IEEE International Workshop on Biometrics and Forensics IWBF2022 Best Paper Award
  - Imen Ben Amor, Jean-François Bonastre: Abstract BA-LR , European Academy of Forensic Science Conference EAFS 2022.
  - Imen Ben Amor, Jean-François Bonastre: BA-LR : une approche transparente de comparaison de voix en criminalistique , Journées d'Études sur la Parole JEP2022.
- **Paul-Gauthier Noé**
  - Thèse « Représentation de la preuve pour le respect de la vie privée : inférence bayésienne, preuve compositionnelle et calibration », 26/04/2023 (https://theses.fr/2023AVIG0113) – Prix de thèse AFCP 2024
  - Noé, P. G., Miao, X., Wang, X., Yamagishi, J., Bonastre, J. F., & Matrouf, D. (2023, June). Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
  - Noé, P. G., Nautsch, A., Matrouf, D., Bousquet, P. M., & Bonastre, J. F. (2022, June). Faire le pont entre l'observation et la preuve: Application au respect de la vie privée. In *Journées d'Etudes sur la Parole-JEP2022*.
  - Noé, P. G., Nautsch, A., Matrouf, D., Bousquet, P. M., & Bonastre, J. F. (2022, May). A bridge between features and evidence for binary attribute-driven perfect privacy. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3094-3098). IEEE.

Reconnaissance du locuteur ou « Biométrie » vocale

Et tous ces systèmes sont parfaitement au point ?

Oui tous ces systèmes sont fiables avec quand même quelques petites réserves. La voix par exemple change avec l'âge ou peut être atteinte, affectée par une maladie. On peut également imaginer que tout système biométrique comme les autres peut être piraté.

Reconnaissance du locuteur ou « Biométrie » vocale

• « Remplacer un mot de passe, sécuriser un paiement… » ✓	• Oui mais « vérification »
• « La voix est unique un peu comme une empreinte digitale » ✗ ✗	• Une très mauvaise analogie, prendre plutôt « signature manuscrite »
• « Plus d'une centaine de caractéristiques sont potentiellement identifiables » ~	• Potentiellement ? Décrites où ? 100 ?
• « La biométrie vocale permet donc de reconnaitre l'empreinte vocale d'une personne » ✗ ✗	• **L'« empreinte vocale» n'existe pas !!** (et c'est très largement admis)
• « Certaines de ces caractéristiques sont physiques, donc quasi impossibles à reproduire » ✗ ✗	• Traces des caractéristiques physiques ! Pourquoi physique = impossible à reproduire ?
• [caractéristiques] « Comportementales » ✓	• Est-ce de la biométrie ?
• « Il suffit de dire une phrase à haute voix pour être formellement identifié et le système est réputé très fiable » ✗	• La fiabilité est fortement questionnée Et reconnue comme modérée (mais la voix a d'autres avantages !!!)

La pression marketing pour l'IA… Exemple de "Your Voice is unique"…

- Système d'identification par la voix mis en place par HSBC en 2016 et "disant que chaque voix est unique et représentée par 100 caractéristiques"
- Les clients accèdent à leur compte en donnant simplement les identifiants (num client/compte) et la phrase "My voice is my password".

Dan Simmons (19/05/2017, BBC) **BBC fools HSBC voice recognition security system**
http://www.bbc.com/news/technology-39965545