Deep modeling based on voice attributes for explainable speaker recognition : application in the forensic domain

Imen Ben Amor

THÈSE DE DOCTORAT D'AVIGNON UNIVERSITÉ

École Doctorale n°536
Agrosciences et Sciences

Mention de doctorat :
Informatique

Laboratoire Informatique d'Avignon

Présentée par
Imen Ben-Amor

Deep modeling based on voice attributes for explainable speaker recognition

Application in the forensic domain

Soutenue publiquement le 25/04/2024 devant le jury composé de :

Alessandro Vinciarelli	Professeur	University of Glasgow	Rapporteur
Tomi Kinnunen	Professeur	University of Eastern Finland	Rapporteur
Didier Meuwly	Professeur	University of Twente	Examinateur
Tanja Schultz	Professeure	University Bremen	Examinatrice
Corinne Fredouille	Professeure	Avignon Université	Examinatrice
Jean-François Bonastre	Professeur	INRIA	Directeur de thèse

Contents

Résumé xi

Abstract xiii

List of acronyms xiv

List of figures xx

List of tables xxii

1 Introduction 1

I Literature Review 7

2 DNN-based automatic speaker recognition systems 9
- 2.1 Introduction . . . 10
- 2.2 A look back at statistical models . . . 10
- 2.3 DNN-based ASpR framework . . . 11
  - 2.3.1 Feature extraction: hand-crafted features . . . 12
  - 2.3.2 Feature extraction: self-supervised features . . . 14
  - 2.3.3 DNN speaker model . . . 16
  - 2.3.4 Scoring . . . 23
- 2.4 Evaluation protocols and metrics . . . 25
  - 2.4.1 Equal Error Rate . . . 26
  - 2.4.2 Detection Cost Function . . . 26
  - 2.4.3 DET curve . . . 27
- 2.5 Summary . . . 27

3 Forensic application of automatic speaker recognition 29
- 3.1 Forensic Automatic Speaker Recognition . . . 30
- 3.2 Bayes paradigm assessing the value of evidence . . . 31
  - 3.2.1 From frequentist to Bayesian approach . . . 31
  - 3.2.2 The Bayesian interpretation and the court . . . 32
- 3.3 Centrality of likelihood ratio . . . 33
  - 3.3.1 LR interpretation . . . 34
  - 3.3.2 LR estimation from similarity scores . . . 34
  - 3.3.3 Calibration of LR into well-calibrated LR . . . 36
- 3.4 On the use of automatic methods in forensic science . . . 36
  - 3.4.1 Acceptability by the court . . . 37
  - 3.4.2 A help or a burden for forensic scientist? . . . 38
  - 3.4.3 Requirement for explanations in forensic science . . . 39
- 3.5 Summary . . . 40

4 Interpretability and explainability in AI **41**
4.1 Introduction . . . 42
4.2 Context of explainable AI . . . 42
4.2.1 Awareness of explainability . . . 43
4.2.2 The need of explainability . . . 43
4.3 Terminology . . . 45
4.3.1 Interpretability . . . 45
4.3.2 Explainability . . . 46
4.3.3 Other related concepts . . . 47
4.4 Literature review: Taxonomy . . . 47
4.4.1 Local Vs. Global . . . 48
4.4.2 Inherently Vs. Intrinsic Vs. Post-hoc . . . 48
4.4.3 Model-specific Vs. Model-Agnostic . . . 49
4.5 XAI for speech system decision . . . 53
4.6 Challenges . . . 54
4.7 Conclusion . . . 54

II Proposed solution and contributions **57**

5 Inspiration and proposed methodology **59**
5.1 Introduction . . . 60
5.2 Drawing inspiration from DNA: Caution is required! . . . 60
5.2.1 Biological traits information . . . 61
5.2.2 DNA individualisation . . . 62
5.2.3 Misleading phenomena . . . 64
5.2.4 Speech is not DNA! . . . 65
5.3 Proposed methodology . . . 66
5.3.1 Assumptions . . . 66
5.3.2 An overview of the approach . . . 67
5.4 Positioning in the interpretability/explainability dilemma . . . 68
5.5 Conclusion . . . 70

6 Step 1: Binary and Attribute-based modelling of speaker embeddings **71**
6.1 Introduction . . . 72
6.2 Related work: Binary speaker embedding . . . 72
6.3 Binary-Attribute-based modelling . . . 73
6.3.1 Requirements . . . 73
6.3.2 Proposed model: ResNet with thresholding . . . 74
6.4 Experiments and results . . . 76
6.4.1 Setup . . . 77
6.4.2 BA-vectors analysis . . . 78
6.4.3 Measuring attributes correlation and dependence . . . 79
6.4.4 Speaker recognition performance . . . 81
6.5 Discussion . . . 82

7 Step 2: Binary-Attribute-based likelihood ratio estimation 85
7.1 Introduction . . . 86
7.2 The core concept of BA-LR . . . 86
7.3 Estimation of behavioral parameters . . . 89
7.3.1 Typicality . . . 90
7.3.2 Drop-out . . . 91
7.3.3 Drop-in . . . 92
7.4 Estimation of likelihood ratios . . . 92
7.4.1 Attribute LR estimation . . . 93
7.4.2 Global LR estimation . . . 96
7.5 Analyses and evaluation of speaker recognition performance . . . 97
7.5.1 Data sets and protocols . . . 97
7.5.2 Analyses of behavioral parameters . . . 98
7.5.3 Speaker recognition performance and generalization ability . . . 100
7.6 Explainability of the LR . . . 100
7.6.1 Distribution of attribute-LLR values . . . 101
7.6.2 Shapley-like explanations . . . 102
7.7 Discussion and perspectives . . . 104

8 Step 3: Attribute explainability 109
8.1 Introduction . . . 110
8.2 The three-world explainability method . . . 111
8.3 Utterance-level: attribute phonetic description . . . 112
8.3.1 Methodology . . . 112
8.3.2 Inherently interpretable classifier . . . 114
8.3.3 Statistical test . . . 115
8.4 Experiments and results . . . 116
8.4.1 Setup . . . 116
8.4.2 Discrimination ability of the attribute model . . . 118
8.4.3 Attribute explainability in terms of phonetics . . . 119
8.5 Frame-level: attribute phonemic and temporal description . . . 123
8.5.1 Attribute related frame-level information . . . 123
8.5.2 Attribute explainability in terms of phonemes . . . 126
8.5.3 Attribute explainability in terms of localized temporal information 129
8.6 Discussion and perspectives . . . 132

III Improvements and application of our approach in forensic context 135

9 Calibration and application of BA-LR on forensically realistic database 137
9.1 Introduction . . . 138
9.2 Global calibration . . . 138
9.3 Selective fusion of attribute LLRs . . . 140
9.3.1 Weighted fusion of attributes . . . 140
9.3.2 Selection of attributes . . . 141
9.4 Experimental protocol . . . 141

9.4.1 Database description . . . 141
9.4.2 Experiments setup . . . 143
9.5 Calibration and speaker recognition performance . . . 146
9.6 Discussion . . . 149

10 Attribute-based binary auto-encoder 151
10.1 Introduction . . . 152
10.2 SPINE: Sparse binarized speaker representation . . . 152
10.2.1 Why sparsity? . . . 153
10.2.2 SPINE model . . . 153
10.3 BAE: attribute-based Binary Auto-Encoder . . . 155
10.3.1 Binary auto-encoder model . . . 155
10.3.2 Proposed attribute-oriented loss . . . 156
10.4 Experimental protocol and analyses . . . 158
10.4.1 Setup . . . 159
10.4.2 Compliance with attribute-based criteria . . . 160
10.5 Speaker recognition evaluation . . . 162
10.5.1 Using cosine similarity scores . . . 162
10.5.2 Application of BA-LR on BAE-vectors . . . 163
10.6 Discussion and perspectives . . . 165

11 Conclusion and perspectives 167

IV Appendices 173

A Extracts from Case text and judicial articles 175
A.1 Courts positions in use cases . . . 175
A.1.1 Over reliance position . . . 175
A.1.2 No trust position . . . 176
A.1.3 Reasonable reliance position . . . 176
A.2 Judicial articles . . . 177
A.2.1 Article 6 from the European Court of Human Rights . . . 177
A.2.2 Article 149 from The Belgian Constitution . . . 177
A.3 GDPR articles . . . 178
A.3.1 Article 15: Right of access by the data subject . . . 178
A.3.2 Article 22: Automated individual decision-making . . . 178
A.4 AI act: Recital 38 . . . 179
A.5 The Equal Credit Opportunity Act . . . 179

B Appendix to Step 2 181

C Extra analyses and results of Step 3 183
C.1 Details of the BA models . . . 183
C.2 Correlation between attributes in terms of frames . . . 184
C.3 Alignment of MegaFrames with phonemes for an attribute . . . 185

D Extra results and analyses of BAE and SPINE vectors **189**
D.1 Extra explainability analyses of BA-LR scoring on BAE system 189
D.2 An additional investigation of Step 3 on SPINE-vectors in JSALT workshop 190
D.2.1 Evaluation of SPINE representations 190
D.2.2 Explainability of SPINE representations 191
D.2.3 Discussion & perspectives 196

E Extra results for BA-LR calibration **199**

إلى أمي و أبي، مصدر قوتي و إلهامي

إلى صديقي و حبيبي و زوجي؛ زياد

Acknowledgment

I would like to thank my defense committee members, namely Pr.*Alessandro Vinciarelli*, Pr.*Tomi Kinnunen*, Pr.*Didier Meuwly*, Pr.*Tanja Schultz*, and Pr.*Corinne Fredouille*, for generously dedicating their time and expertise to evaluate this manuscript. Their insightful feedback has enriched this dissertation, enhancing its rigor and relevance to a remarkable extent. A heartfelt appreciation goes to Pr.*Didier Meuwly* for his unconditional support during this thesis. His advice and encouragement have been incredibly helpful, motivating me to believe in this work and move forward.

To my dear supervisor, *Jean-François Bonastre*, thank you for being my inspiration in this work. Thank you for every time you told me *"Imen, there is nothing you cannot do!"*. These words have meant the world to me and have fueled my motivation. Your belief in my abilities has been a guiding light throughout this journey. Thank you for embracing my weaknesses and guiding me towards improvement, allowing me to become the best version of myself. This work was never easy, nor was it evident, but "WE" did it. You were the source of my bravery. I am truly grateful for your time, support, guidance, advice, and, most importantly, for believing in me since the first day I arrived in the lab.

To all the wonderful people I've encountered at the LIA laboratory as well as at CERI, it has been a pleasure knowing you. Thank you dear friends and colleagues for being a part of this journey. A special thanks for *Salima* for being always here when I needed. To *Aran*, my office-mate during all the three years, thank you for the very interesting discussions, the brainstorming and for your attentive listening and advises. To *Sarkis*, *Ahmed*, *Virgile*, *Chaimae*, *Othman*, *Benjamin* and other permanent and non permanent colleagues, thank you for your presence in my defense and for sharing this special moment with me. A heartfelt thanks to Dr.*Driss Matrouf* for his support, encouraging words, and contagious good humor that never fails to change my mood. Thank you for always being available and for the insightful exchanges of ideas. A special thanks to Dr.*Teva Merlin* for generously dedicating time to resolve all my logistical issues and for providing constructive feedback on my presentations. I would like also to express my gratitude to the sponsors of the Avignon Chair of Artificial Intelligence, *LIAvignon*, for funding this thesis.

Many thanks to the people I met at conferences, at JSALT workshop, and during my visit to the Netherlands Forensic Institute (NFI). Each of them has directly or indirectly contributed to this work by offering advice, sharing ideas, engaging in discussions, and providing resources. A special thanks to *David Van Der Vloed*, a forensic practitioner at NFI, who dedicated his time and effort to make my journey at NFI easier and smoother.

****

To my parents, *Salwa* and *Mohamed*, who couldn't be present at my thesis defense, you were always in my heart, guiding me every step of the way. Thank you for supporting all my choices and believing in me, even though it was challenging for both of you. Your sacrifices have not gone unnoticed. You have been my source of strength in the weakest moments, my guiding light in the darkest moments, and my motivation to fight until the end. Having reached this significant milestone, I am immensely proud to dedicate all my success and achievements to you.

****

To my beloved husband, *Zied*, you have been my rock, my confidant, my love, my friend, my partner and my everything. Your unconditional support, understanding, and sacrifices have been instrumental in completing this work in such a short time. Thank you for your patience, empathy, assistance, and, above all, for being an attentive listener. I am deeply grateful for the immense effort you put into traveling for six hours twice a week from Bordeaux to Avignon just to spend one weekend with me. Your understanding when I am unable to cook or clean the house, or when I am feeling overwhelmed, means the world to me. Thank you for being there for me during my toughest moments. Thank you for your unconditional love and support. Your presence has made all the difference.

****

To my sisters, aside from your never-ending requests for clothes, your support and belief in me have been my guiding light. Not to forget that thanks to you also, my bank account is forever in the red :D. Dear *Amal*, thank you for being the kindest sister ever, my problem solver, my shelter and my rock through it all. Dear *Ines*, thanks for always teasing me and bringing a smile to my face even when I'm feeling down. Dear *Eya*, thank you for always being proud of me. I have no doubt that an amazing future awaits you. Just keep the good work. To the rest of my family, near and far, who have encouraged and supported me throughout this journey.

****

In loving memory of the kindest soul, my Grandmother *Asia*, who passed away during my second year of thesis, a departure that shattered me. You will forever reside in my heart. May your spirit find eternal peace.

****

To my friends, who have always been there to lift me up when I'm feeling down, to listen to me when I'm feeling low, who have cared for me, and who have always believed in me.

To *Sondes*, who stood by my side throughout this thesis journey. Your constant presence, support, and kindness have meant the world to me. Thank you for being such a wonderful and caring friend. I appreciate your attentive listening and our fruitful discussions, which have helped shape and organize my thoughts and ideas. You have not only acted as a second supervisor in my thesis but also as a close friend who always makes me feel comfortable and understood when we talk. Thank you so much for believing in me. I will always miss our coffee breaks and our work, life, love discussions ;).

To my baby girls, *"Mes amours"*, my childhood friends since the age of 16. Despite the miles between us, distance has never hindered our bond. You reside always in my heart, and I'll forever cherish the unwavering support you have shown from our college days to my PhD journey. To *Jihen*, I am truly grateful for your travelling to Avignon to support me during my defense and to share this special moment with me.

Thank you for your care and for being there for me, especially during times when I felt depressed and burned out. You have been the sweetest psychologist to me, providing comfort and understanding when I needed it most. To *Marwa*, you were always my lovely advisor, thank you for having that magical touch on my heart which makes me feel better. To *Wiem*, my partner in crime during our craziest teenage moments, my study buddy through failures and successes alike. Despite the distance and the paths life has taken us on, you will always be my soulmate. You always hold a special place in my heart. To Islem, who has always been incredibly caring and understanding. You have always encouraged me, and your genuine happiness for my successes means a lot for me. Thank you for your infinitely kind heart.

****

Last but certainly not least, I want to extend my deepest gratitude to myself for persevering until the end. Thank you for bravely navigating through my moments of depression, my darkest hours, bouts of sadness, and the whirlwind of emotions. I am truly grateful for your resilience in facing exhausting stress, fatigue, sleepless nights, and the myriad of challenges. Your ability to handle copious amounts of coffee consumed speaks volumes. Thank you for bravely confronting negative thoughts, feelings of hopelessness, and moments of burnout. Above all, I want to acknowledge that pushing you to your limits has revealed my inner strength, capabilities, and a better version of myself that I never knew existed. I am so proud of the person I have become.

Résumé

La Reconnaissance Automatique du Locuteur (RAL) a été intégrée dans différentes applications, allant de la sécurisation des accès ou l'identification en criminalistique. Son objectif est de déterminer automatiquement si deux échantillons vocaux proviennent du même locuteur. Les systèmes de RAL reposent principalement sur des réseaux neuronaux (DNN) complexes et présentent leurs résultats par une seule valeur. Malgré leurs performances élevées, ils sont incapables de fournir des informations sur la nature des représentations vocales utilisées, leur encodage et leur influence sur la prise de décision. Ce manque de transparence pose d'importants défis pour aborder les préoccupations éthiques et légales, en particulier dans des applications à haut risque telles que la comparison de voix criminalistique. Cette thèse introduit une approche en trois étapes basée sur l'apprentissage profond, conçue pour fournir des résultats de RAL interprétables et explicables.

Dans la première étape, nous représentons un extrait vocal par la présence ou l'absence d'un ensemble d'attributs vocaux, partagés entre des groupes de locuteurs et sélectionnés pour être discriminants du point de vue locuteur. Cette information est encodée par un vecteur binaire où un coefficient égal à 1 indique la présence de l'attribut correspondant dans l'extrait vocal et 0 son absence. Cette représentation permet d'apporter de l'interprétabilité, tout en offrant un niveau de performance proche de celui des systèmes état de l'art (SOTA) de RAL.

La deuxième étape s'intéresse au calcul explicite du score de RAL, représenté ici par un rapport de vraisemblance (LR). Nous proposons pour cela une méthode denommée BA-LR qui décompose le processus de calcul en sous-processus, chacun dédié à un attribut. Un LR d'attribut, est estimé pour chaque attribut en utilisant uniquement la présence ou l'absence de celui-ci et sa description, définie par trois paramètres comportementaux explicites. Le LR final est calculé comme le produit des LR d'attribut, en supposant leur indépendance. Cette estimation permet un calcul transparent du LR, associé à des explications détaillées sur la contribution de chaque attribut à la valeur finale du LR, à même d'aider les utilisateurs, tels que les juges, dans leur prise de décision.

La troisième étape est dédiée à la découverte de la nature des attributs. Nous proposons une description automatique des attributs en informations acoustiques, phonétiques et phonémiques à l'aide de différentes méthodes d'explicabilité. Les explications obtenues permettent de mieux appréhender les attributs de la voix utilisés en RAL et offrent des perspectives pour les phonéticiens.

Pour valider l'efficacité de notre approche en criminalistique, nous l'avons évaluée à l'aide d'une base de données spécifique à ce domaine. Nous avons défini pour cela une approche de calibration adaptée aux domaine. Les résultats démontrent la robustesse et la capacité de généralisation de BA-LR dans un contexte criminalistique.

Les différentes contributions de cette thèse ouvrent une nouvelle perspective en termes d'explicabilité en RAL, en proposant d'accompagner l'inférence, le LR, par les explications nécessaires à une prise de décision transparente, avec un niveau de performance comparable aux systèmes SOTA. En criminalistique, notre approche semble prometteuse, facilitant la compréhension des éléments de décision par les experts et leur prise en compte par la cour. Elle offre également aux phonéticiens un outil pour mieux comprendre les informations vocales. Toutefois, ces résultats encourageants doivent

être approfondis avec une variété de cas d'utilisation avant d'être appliqués dans des contextes réels en criminalistique, en respectant le "devoir de précaution" propre à ce domaine.

Abstract

Automatic speaker recognition (ASpR) has been integrated into critical applications, ranging from customised assistant services to security systems and forensic investigations. It aims to automatically determine whether two voice samples originate from the same speaker. These systems primarily rely on complex deep neural networks (DNN) and present their results by a single value. Despite the high performance demonstrated by DNN-based ASpR systems, they struggle to provide transparent insights into the nature of speech representations, its encoding, and its use in decision-making process. This lack of transparency presents serious challenges in addressing ethical and legal concerns, particularly in high-stakes applications such as forensics.
This thesis introduces a three-step methodology based on deep learning, designed to provide interpretable and explainable ASpR results.

In the first step, we represent a speech extract by the presence or absence of a set of speech attributes, shared among groups of speakers and selected to be speaker discriminant. This information is encoded by a binary vector where a coefficient equal to 1 represent the presence of the corresponding attribute in the speech extract and 0 its absence. This binary and attribute-based modelling facilitates interpretability and allows for a better handle of the speech information. The results show that the obtained representations are more interpretable and offer a level of performance close to that of State-Of-The-Art (SOTA) ASpR.

In the second step, the goal is to ensure transparent computation of the likelihood ratio (LR), thereby facilitating a more informed assessment of the value of speech evidence in a courtroom setting. We therefore propose the *Binary-Attribute-based LR* (BA-LR) framework, that breaks down the scoring process into independent subprocesses, each dedicated to an attribute. An attribute-LR is a LR estimated using only the presence or absence of the attribute and its description, defined by three explicit behavioral parameters. The final LR is calculated as the product of the attribute-LRs, assuming independence between them. This framework enables transparent LR computation and a clearer understanding of the value of evidence. It also provides detailed explanations of the contribution of each attribute's information to the final LR value, aiding juries and judges in decision-making.

In the third step, we conduct a discovery of the nature of attributes. This investigation employs statistical techniques, surrogate models as well as backpropagation and alignment strategies to provide a description of attributes in terms of acoustic, phonetic and phonemic information. The obtained explanations serve as a valuable tool for phoneticians to interpret the contributing attributes to a given LR.

Additionally, our three-step approach is validated through the application of BA-LR on a forensically realistic dataset. In such context, we apply a Logistic Regression model to handle the mismatch between the training conditions and a real-world scenarios. Results demonstrate the robustness and the generalisation ability of BA-LR in a forensic context.

Overall, this thesis opens a new perspective on explainable ASpR, by proposing a solution for a transparent decision making, with a level of performance comparable to SOTA systems. Our approach shows promise in offering forensic practitioners and the court insights into the value of evidence while also serving as a discovery tool for phoneticians helping them better understand and interpret speech information. As

always in the field of forensics, these encouraging results require further evaluation through additional studies before being applied in real-world situations.

Acronyms

AI Artificial intelligence.

ASpR Automatic Speaker Recognition.

BA Binary Attribute.

BA-LR Binary-attribute-based Likelihood ratio.

BAE attribute-based Binary Auto-encoder.

C_{llr} Cost of Log Likelihood Ratio.

CNN Convolution Neural Networks.

DNA Deoxyribonucleic acid.

DNN Deep Neural Network.

EER Equal Error Rate.

FASpR Forensic Automatic Speaker Recognition.

FSpR Forensic Speaker Recognition.

GDPR General Data Protection Regulation.

GMM-UBM Gaussian Mixture Model-based Universal Background Model.

LLD Low Level Descriptors.

LLR Log Likelihood Ratio.

LR Likelihood Ratio.

LRP Layer-wise Relevance Propagation.

MF MegaFrames.

MFCC Mel-Frequency spaced Cepstral Coefficient.

MSE Mean Squared Error.

NFI Netherlands Forensic institute.

PCR Polymerase Chain Reaction.

PLDA Probabilistic Linear Discriminant Analysis.

ResNet Residual Neural Network.

SLDA Stepwise Linear Discriminant Analysis.

SPINE SParse Interpretable Neural Embeddings.

SSL Self Supervised Learning.

TDNN Time Delay Neural networks.

UBM Universal Background Model.

VAD Voice Activity Detection.

VQ Vector Quantization.

XAI Explainable Artificial Intelligence.

List of Figures

1.1 An overview of our three-step approach . . . 3

2.1 The VQ codebook training[20] . . . 10
2.2 DNN-based ASpR framework . . . 12
2.3 Preprocessing and feature extraction block diagram . . . 12
2.4 The architecture of wav2vec 2.0 model . . . 15
2.5 The architecture of WavLM model [38] . . . 16
2.6 The components of a DNN speaker model . . . 17
2.7 TDNN computation diagram [43] . . . 18
2.8 TDNN architecture configuration for x-vectors [47] . . . 18
2.9 Identity Shortcut connection [56] . . . 19
2.10 ResNet configuration for x-vectors [57] . . . 20
2.11 Structure of the SE-Res2Block of the ECAPA-TDNN architecture [71] . 20
2.12 The architecture of ECAPA-TDNN [71] . . . 21
2.13 Relationship between FAR, FRR and EER . . . 26
2.14 Plot of DET Curves for a speaker recognition evaluation [96] . . . 27

3.1 Speaker recognition Vs. Forensic Speaker Recognition . . . 30
3.2 Our proposed categorization of the different positions of the court regarding the use of automatic methods . . . 38

4.1 Google trends for research interest of the terms "interpretability" and "explainability" from 2004 until September 2023 . . . 43
4.2 Our proposed explainability and Interpretability taxonomy . . . 48
4.3 Inherently Vs. Intrinsically interpretable/explainable models . . . 48
4.4 Post-hoc explainability methods . . . 49
4.5 Train a surrogate model on input and output of black box model . . . 50
4.6 SHAP explanations of a black box model . . . 52

5.1 Process of alleles information extraction from DNA . . . 60
5.2 Quantitative and qualitative biological traits . . . 61
5.3 DNA analogy to voice characteristic . . . 62
5.4 DNA individualisation . . . 63
5.5 An illustration of assumptions of the ideal representation . . . 66
5.6 Positioning of our approach in the interpretability/explainability dilemma 69

6.1 ResNet training with thresholding function . . . 74
6.2 ReLU and Softplus activation functions, along with post-hoc thresholding functions. . . . 76
6.3 Distribution of 0s in the BA-vectors . . . 79
6.4 Pearson correlation between neurons activation . . . 80
6.5 Distribution of correlation values in boxplots . . . 80
6.6 Mutual information between binary attributes . . . 81

7.1 An illustration of the assessment of the value of speech evidence using ASpR system . . . 87

7.2 An illustration of an interpretable assessment of the value of speech evidence using ASpR system . . . 88
7.3 A dreamlike interpretable likelihood ratio calculation . . . 88
7.4 Global likelihood ratio estimation using BA-LR . . . 93
7.5 The number of speakers as a function of the number of present attributes in their profiles . . . 98
7.6 Distributions of attributes typicality and drop-out values . . . 99
7.7 Estimation of the optimal value of Din on the training data . . . 99
7.8 Distribution of attribute LLRs values for 00, 11 and 01|10 cases, computed on VoxCeleb1 comparison pairs . . . 101
7.9 Relationship between attribute-LLRs of DNA-inspired and Speech-adapted versions . . . 102
7.10 Attributes contributions to the final LLR for a target (above) and a non-target (below) comparison pairs. . . . 103
7.11 Average contribution of attributes among all pairs Vs. their behavioral parameters, typicality & drop-out . . . 104
7.12 Overview of Step 1 & 2 of our approach with interpretability and explainability aspects . . . 105

8.1 The three-world illustration of our proposed methodology . . . 111
8.2 Methodology of an attribute explainability following (a), (b), and (c) sub-steps, as applied to each attribute . . . 113
8.3 Application of the Decisiontree classifier for Attribute BA_i with TreeExplainer from SHAP . . . 115
8.4 Accuracy of attribute models on Vox2 (train) and Vox1 (test) along with their associated typicality values. . . . 118
8.5 Wilk's Lambda values as a function of the 65 selected variables for attribute BA_9, with a closer look on the first 10 variables . . . 120
8.6 Evolution of accuracy of sub-models of BA_9, each trained with incremental number of descriptive variables: from most to least contributive 120
8.7 Descriptive variables contributions to the BA_9 model, grouped by families 121
8.8 Heatmap illustrating the contribution of variable families to attributes grouped into clusters . . . 122
8.9 DNN architecture of the BA-extractor . . . 124
8.10 Modified DNN architecture for frame-level information extraction . . . 124
8.11 A closer look into the ResNet extractor illustrating the relationship between the input frames and the output MegaFrames through ResNet blocks. . . . 125
8.12 An example of attribute alignment with input frames through selected MegaFrames. . . . 126
8.13 Normalized activations of MFs to BA_{11}=1 in a portion of 0.8s of a speech utterance of 5s, aligned with phonemes and classes of phonemes . . . 127
8.14 Occurrence of each class of phonemes, clustered per BAs . . . 128
8.15 Occurrence of each phoneme in its phonetic class . . . 129
8.16 Occurrence of all phonemes together, clustered by BA . . . 130
8.17 Distribution of the mean and std values per descriptor across all BAs . 131

9.1 Global calibration of the final LLR using univariate logistic regression . 139
9.2 A weighted fusion of attribute-LLRs using a multivariate logistic regression model 140
9.3 Description of experimental protocol using calibration and fusion approaches 144
9.4 An example of EER and Cllr_{cal} evolution using BA-LR, along with the number of attributes for the optimal fold 148

10.1 SPINE architecture adapted to speech 153
10.2 Architecture of the attribute-based Binary Auto-Encoder 156
10.3 A toy example illustrating the relationship between utterances activations and speakers profiles used to compute the typicality 157
10.4 Illustration of the sparsity loss computation during training using the latent space representation before binarization 157
10.5 Snapshot of the losses evolution during the training ofBAE model . . . 160
10.6 Distributions of Pearson correlation values between dimensions 161
10.7 Sorted typicality values across dimensions of the three binary vectors. . 161
10.8 Relationship between typicality and drop-out of BA-vectors and BAE vectors 163
10.9 Search for the optimal drop-in value for each version of BA-LR using the train set of BAE vectors 164

B.1 All possible combinations of observed and real (i.e. actual) state for Speech-based version of BA-LR 181

C.1 A heatmap illustrating the binarization of the Softplus-matrix A for a given utterance, with clustering of BAs in terms of MegaFrames using Jaccard distance. 184
C.2 Correlation between attributes in terms of MegaFrames contributing to 100% 185
C.3 MegaFrames activations to BA_{10}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes 185
C.4 MegaFrames activations to BA_{11}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes 186
C.5 MegaFrames activations to BA_{12}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes 186
C.6 MegaFrames activations to BA_{17}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes 186
C.7 MegaFrames activations to BA_{23}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes 187

D.1 Distribution of attribute LLRs using BAE-vectors 189
D.2 Relationship between behavioral parameters and the contribution of attribute LLRs using BAE-vectors 190
D.3 EER% using SPINE-vectors on VoxCeleb1 191
D.4 An overview of the explainability schema for SPINE representations . . 191
D.5 Most important features for gender detection selected using three methods 193

D.6 Phonetic description of gender-specific features, grouped by families of descriptors 195
D.7 Phonetic description of emotion-specific features, grouped by families of descriptors 195
D.8 Pearson correlation between OpenSmile eGeMAPs descriptors 198

List of Tables

2.1 Comparison of speaker recognition performance in terms of EER (defined later in this chapter) between pretrained Wav2vec and WavLM [38] on SUPERB Benchmark [40] . . . 16
2.2 Comparison of speaker recognition performance in terms of EER between different model architectures on VoxCeleb1-O . . . 22
2.3 Comparison of scoring backends of ResNet model using EER for two test datasets . . . 25

4.1 Axioms definition of Shapley values . . . 52

6.1 Data set and protocol description . . . 77
6.2 Performance comparison in terms of EER and C_{llr} of the three systems on VoxCeleb1 using cosine similarity. . . . 81

7.1 Description of data sets . . . 97
7.2 Speaker recognition performance in terms of EER and $C_{llrmin/act}$ . . . 100
7.3 Details about the most contributing attributes for two speech pairs, a target pair and a non-target pair . . . 103

8.1 Phonemes and corresponding classes based on MFA chart . . . 127

9.1 Recording devices description . . . 142
9.2 Sessions description . . . 143
9.3 Experiment data description . . . 144
9.4 Description of Dev and test sets of the best fold for each experiment . . . 145
9.5 Speaker recognition performance of BA-LR Speech-based version on Test sets before and after selective fusion . . . 146
9.6 Speaker recognition performance of X-vectors on all comparison pairs. These results are indicative only, as they are calculated based on all comparison pairs corresponding to the combination of (device, sessions). 147
9.7 $\text{Cllr}_{min/act}$ computed with BA-LR before (Non-Calibrated) and after (Calibrated) applying calibration and fusion approaches (results for the best fold) . . . 147

10.1 SPINE configuration for three systems . . . 159
10.2 Speaker recognition performance of the three systems on VoxCeleb1 in terms of EER using cosine similarity scoring . . . 162
10.3 Speaker recognition performance of BAE system and BA-extractor on VoxCeleb1 in terms of EER and $Cllr_{min/act}$ . . . 164

C.1 Further details about the number of speech extracts selected for the first 16 BAs (out of 205) for train and test datasets. . . . 183

D.1 Number of samples per class for emotion and gender detection . . . 192
D.2 Evaluation of performance in emotion and gender detection tasks . . . 192
D.3 Families of descriptors . . . 194

E.1 Comparison of speaker recognition performance of BA-LR Speech-based and DNA-inspired versions on Test sets before and after fusion 199
E.2 Comparison of $\text{Cllr}_{min/act}$ between DNA-inspired and Speech-based versions of BA-LR before (Non-Calibrated) and after (Calibrated) applying calibration and fusion approaches 199

Chapter 1

Introduction

The richness of human voice is an extraordinary phenomenon, that goes far beyond simple language transmission. It conveys not only words, but a huge amount of details about the speaker. Each person's voice is a combination of elements such as pitch, tone, rhythm and resonance. This combination holds a remarkable potential to uncover the speaker's identity, decode his/her emotional state, and even unveil his demographic trait. In human-to-human interactions, our brain's ability to recognize and identify individuals by their vocal characteristics is an innate skill, often taken for granted in our everyday lives. Consider the moment when we pick up the phone to engage in conversation, before any exchange, we subconsciously start by confirming the identity of the person on the other end. We recognize our loved ones, friends, and people we are familiar with by the sound of their voices. This intuitive process, known as *naïve speaker recognition* [1], is a testament to our remarkable ability to discern familiar voices from the crowd.

In our modern technology-driven world, speaker recognition applications extend far beyond everyday interactions. Speaker recognition has transitioned into a fully automated process where machines carry out the recognition task, known as **Automatic Speaker Recognition (ASpR)**. It finds applications across various fields [2, 1, 3]. For instance, within the domain of audio archives management [4], ASpR technology has been integrated into the workflow of archivists. This technology accesses the wealth of personal information embedded within a speaker's voice in order to provide more efficient organization and structuring of audio content. ASpR systems have also emerged in more sensitive fields such as security and access control to confidential information. These systems operate by either granting or denying access based on voice authentication. This approach is mainly employed in secure facilities and banking applications. Furthermore, ASpR technology finds utility in voice assistants[1]. It serves for both user authentication and personalization such as interacting based on individual preferences and profile.

Another particular application of ASpR systems lies in forensic investigations. Traditionally, forensic experts utilize analysis techniques to extract and determine whether

[1] e.g. Siri, Alexa...

the vocal characteristics of a trace speech sample align with those of a suspect recording. Then, a **Likelihood Ratio (LR)** report is generated, evaluating the value of evidence. This ratio indicates the likelihood that the two voice samples originate from the same individual divided by the likelihood that they belong to different individuals. This entire process is known as *Forensic Voice Comparison*. Recently, the integration of ASpR technology has automated this process, now referred to as **Forensic Automatic Speaker Recognition (FASpR)**. This transition has been mostly driven by the increasing use of cell phones for criminal communications. ASpR systems are now employed to extract information from audio evidence (e.g. telephone intercepts), providing useful insights to the investigation of criminal cases.

Inspired by the widespread adoption of artificial intelligence (AI) in real-world applications, speaker recognition systems have also emerged as beneficiaries of the robustness and the high performance of **Deep Neural Networks (DNN)**. In this scenario, a DNN model is trained to extract **Speaker Embeddings** representing speech data. These embeddings are compared using similarity score, allowing the system to determine whether two audio recordings originate from the same or different speakers. Researchers primarily directed their attention towards enhancing performance through increasingly complex models. However, this complexity comes at the cost of *transparency* and leads to several problems about providing informed decision making. Questions have arisen regarding the fairness and equity of these models [5, 6], such that the system's results lead to decisions that are skewed towards a certain group of individuals with specific voice attributes such as social origins, gender, or age [7]. The biases existing in real-world data might be inherently fed to the model during the learning process. As a result, certain groups may encounter unfair restrictions in accessing and authenticating platforms, while others might become more exposed to potential threats or more susceptible to be identified as criminals. Perhaps one of the most sensitive fields to this issue are forensics and law enforcement applications, where the risk of introducing discrimination bias [5] due to a black box model, is a paramount concern that can cause serious issues. In recent years, concerns have emerged within the criminal justice system regarding the potential for discriminatory practices associated with AI-based algorithms [8]. In 2016, ProPublica's examination [9] of Correctional Offender Management Profiling for Alternative Sanctions tool, COMPAS, revealed that although the algorithm's overall accuracy is similar for both white and black defendants ($\approx 62\%$), the types of errors it makes differ. It tends to categorize black defendants more frequently as high-risk when they are not, while it more often categorizes white defendants as low-risk when they are not. Similarly, another predictive algorithm, PredPol, used by law enforcement, has faced criticism for its potential to disproportionately target low-income, Black communities [10]. Such discrimination raises serious ethical questions regarding the trustfulness on those systems which have decisions over people's lives.

Recent works proposed some recommendations to controlling the model learning and using balanced data to reduce bias [11, 7]. However, with all precautions considered, this bias remains present [11] and constitutes a big challenge to those complex black box models. These issues have fueled a growing demand for increasing transparency in AI models. For instance, the ENFSI Expert Working Group Foren-

sic Speech and Audio Analysis proposed a best practice guide for performing forensic examinations[12], underscoring that, within the context of forensics, simply presenting the output of an automated system is insufficient. To ensure the credibility and ethical integrity of results, additional research are mandatory. This highlights the ethical imperative of ASpR systems **explainability** and **interpretability**[2] to maintain fairness and equitability in its decision-making process.

In this thesis, our primary objective is to address the lack of interpretability and explainability of ASpR systems in general applications with a specific focus on forensic context. Our choice to position ourselves in high-risk scenarios is motivated by the critical significance of interpretability and explainability. In these contexts, relying solely on a single value derived from an ASpR system, namely the LR, proves insufficient. Therefore, the central question guiding this thesis is whether it is possible to enhance confidence in DNN models by providing explanations of the ASpR system output that are easily interpretable in a courtroom setting and comprehensible to forensic experts.

To this end, we propose a three-step approach summarized in Figure.1.1. This approach builds upon existing DNN-based ASpR systems, introducing a novel perspective that prioritizes interpretability and explainability. It is designed in such a way that each step enhances the overall level of interpretability and explainability.

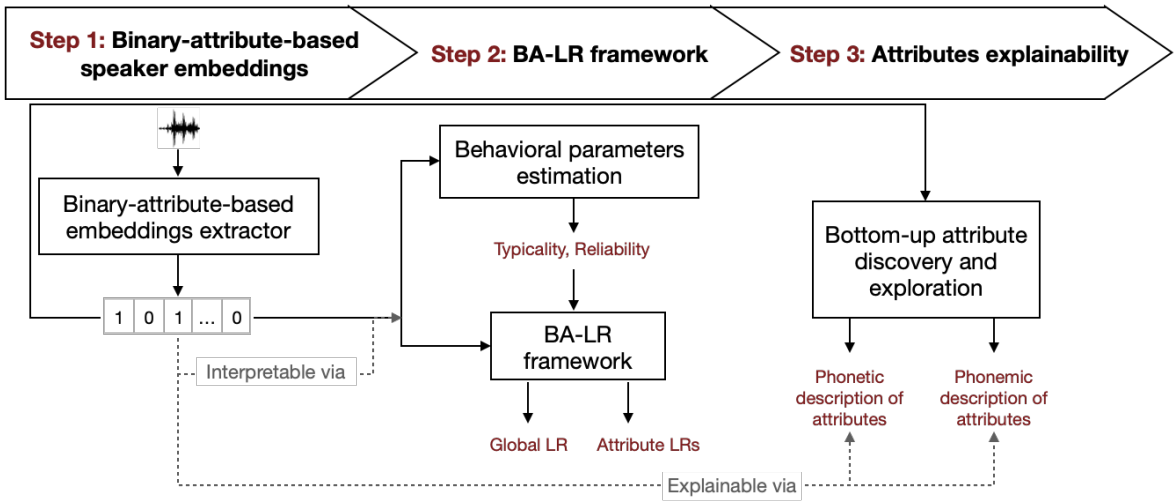


Figure 1.1: An overview of our three-step approach

The first step, namely **Step1**, aims to incorporate an inherent explainability into the speaker embeddings. It mainly drives them towards a desired representation that is more easily interpretable. This representation is an embedding space where the coefficients are binary and encode the presence or absence of a voice attribute (i.e. On or Off). This representation is denoted as *Binary-attribute-based* embedding. The main idea is to propose a binary and attribute-based modeling of speaker embeddings. This step serves as the foundation of our approach, enabling for a bottom-up discovery of attributes.

[2]These two terms are discussed and defined in further details in Chapter 4

The goal of the second step, **Step 2**, is to address the lack of informative and interpretable scores produced by general ASpR systems. The objective of this step is to simplify the computation of the score for better comprehension in a legal setting and to provide additional explanations regarding the information it encapsulates. For this end, we propose the *Binary-Attribute-based LR estimation (BA-LR)* framework. This framework estimates the LR based on the binary attributes encoded in the embeddings. It firstly estimates the "behavior" of these attributes in terms of typicality (i.e. discriminatory power) and reliability. Then it uses this behavior to calculate a LR per attribute, namely an *attribute LR*. The *global LR* of a speech comparison pair is calculated as the product of all these attribute LRs. This framework enables the comprehension of the individual contribution of each attribute to the LR value through attribute LR values. When combined with behavioral parameters for each attribute, it forms a robust explanation of a given LR value. All these explanations represent a very useful tool for the forensic practitioner to understand the output of DNN models and an understandable framework for the court to take in hands the weight of evidence.

Thus far, the level of provided explainability remains incomplete as the information encoded into attributes has not been yet explored. This leads us to the third step of our approach, namely **Step 3**, consisting in a bottom-up attribute discovery. In this step, we introduce a novel methodology that aims to explore the information encoded within binary attributes of speaker embeddings including a range of acoustic, phonetic, phonemic, and temporal descriptors. This would help to understand the nature of attributes and the specific vocal characteristics encoded in speaker embeddings and contributed to the LR computation.

Overall, this work draws attention to an unexplored area such as the interpretability and explainability of DNN-based ASpR systems in general and more specifically in forensic context. It introduces a different perspective to make ASpR system more interpretable and explainable. This solution represents as a whole a powerful tool for experts to comprehend the output of automatic systems and identify any potential discrimination bias. It provides informative support for the court, thereby facilitating the decision-making process.

This thesis is organized into three main parts. The **first part** includes a literature review and defines fundamental concepts related to this work. In Chapter 2, we present existing works in DNN-based ASpR systems and provide an overview of the entire speaker recognition framework. This overview encompasses key components such as feature extraction, DNN speaker modeling, scoring techniques, and evaluation tools. Chapter 3 focuses on the application of ASpR systems in the forensic context, highlighting the adaptation of ASpR systems to the Bayesian framework and the central role of the LR in the judicial process. Particular attention is given to the challenges associated with the use of ASpR methods in forensic, emphasizing the need for caution and the importance of interpretability in DNN models. Chapter 4 further explores this direction. It underscores the adoption of AI interpretability and explainability methods to address the opaqueness in DNN models, both within high-risk contexts, such as forensics, and more broadly. It clarifies the dilemma in the terminology of interpretability and explainability in the literature and define a taxonomy of the dif-

ferent methods. Building upon these theoretical definitions and fundamental concepts, we present in the **second part** our proposed solution and contributions. Chapter 5 is dedicated to firstly introduce the initial inspiration of our approach and give an overview of the three steps, while positioning each step within the explainability and interpretability dilemma. Following that, we dedicate a distinct chapter for each step of our proposed approach. Chapter 6 is focused on describing Step 1 of our solution, providing further details about the proposed binary-attribute-based modelling. Chapter 7 introduces the core concept of the BA-LR[3] scoring proposed in Step 2. Chapter 8 describes the methodology proposed in Step 3 of our approach. It explains the nature of attributes, with an acoustic and phonetic description of attributes. The **third part** is showcasing a real application of our approach and introducing further refinements and improvements. This part offers different perspectives on the three steps of our approach, underscoring its high potential. Additionally, it introduces supplementary work dedicated to enhance the reliability and validity of our approach. Chapter 9 illustrates the application of BA-LR scoring, namely Step 1 & 2, within a forensic context employing a forensically realistic database. In addition to this application, the chapter introduces an adaptation and an improvement of our approach to suit the specific conditions of forensic data. Chapter 10 proposes an improvement over Step 1. It provides a dedicated binary-attribute-based extractor to extract more accurate binary speaker embeddings. This chapter aims to reinforce the validity and the potential of our approach, thereby paving the way for further exploration in this direction. Finally, we conclude this thesis in chapter 11 by offering a comprehensive summary of the key findings. From a broader perspective, we emphasize the main contributions made by this work and we provide insightful suggestions on potential future directions for expanding upon the concepts explored in this thesis.

[3]BA-LR framework has been recognized with the Best Paper Award in [13]

Personal publications

- Imen Ben-Amor and Jean-François Bonastre. *"BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison"*. In: International Workshop on Biometrics and Forensics (IWBF). 2022. **Best Paper Award**
- Imen Ben-Amor and Jean-François Bonastre. *" BA-LR : une approche transparente de comparaison de voix en criminalistique"*. In: Proc. XXXIVe Journees d'etudes sur la Parole – JEP 2022. 2022, pp. 646–654. doi: 10.21437/JEP.2022-68.
- Imen Ben-Amor and Jean-François Bonastre. *"BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison"*. In: 9th European academy of forensic science conference, EAFS 2022, p. 229.
- Imen Ben-Amor et al. *"Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition"*. In: Proc. INTERSPEECH 2023. 2023, pp. 3207–3211. doi: 10.21437/Interspeech.2023-1648.
- Imen Ben-Amor, Jean-François Bonastre, David Van Der Vloed. *"Forensic speaker recognition with BA-LR: calibration and evaluation on a forensically realistic database"*. In: Odyssey 2024.
- Imen Ben-Amor, Jean-François Bonastre, Salima mdhaffar. *"Extraction of interpretable and shared speaker-specific speech attributes through binary auto-encoder"*. In: Proc. Interspeech 2024.

Other publications

- Anaïs Chanclu, Imen Ben-Amor et.al. *"Automatic Classification of Phonation Types in Spontaneous Speech: Towards a New Workflow for the Characterization of Speakers' Voice Quality"*. In:Proc. INTERSPEECH 2021. doi:10.21437/Interspeech.2021-1765.
- Marie Tahon, Imen Ben Amor et.al. *"Interprétabilité pour l'identification de locuteurs. Retour sur le projet JSALT 2023"*, Journée commune AFIA-TLH / AFCP 2023.

Part I

Literature Review

Chapter 2

DNN-based automatic speaker recognition systems

2.1 Introduction 10
2.2 A look back at statistical models 10
2.3 DNN-based ASpR framework 11
2.3.1 Feature extraction: hand-crafted features 12
2.3.2 Feature extraction: self-supervised features 14
2.3.3 DNN speaker model 16
2.3.4 Scoring 23
2.4 Evaluation protocols and metrics 25
2.4.1 Equal Error Rate 26
2.4.2 Detection Cost Function 26
2.4.3 DET curve 27
2.5 Summary 27

This chapter presents a literature review about automatic speaker recognition models. We provide a comprehensive overview of the entire speaker recognition framework, including key components such as feature extraction, speaker modeling, scoring techniques, and evaluation tools. Finally, we conclude the chapter by discussing the foundational choices of this work with respect to existing literature.

2.1 Introduction

The exploration of automatic speaker recognition (ASpR) dates back to the 1960s [14]. Over the subsequent four decades, many technological advancements promoted the evolution of speaker recognition. In the early 2000s, the Gaussian mixture model-based universal background model (GMM-UBM) was introduced [15]. This approach paved the way for several prominent models such as i-vectors [16] until the emergence of deep learning-based speaker recognition. More recently, driven by the performance and robustness of Deep Neural Networks (DNN) models, various approaches have emerged for ASpR, marking a new era in the field.

In this chapter, we begin with an overview of the historical development of statistical ASpR models predating the emergence of DNN models. We then delve into the fundamental components of the DNN-based ASpR framework. Firstly, we describe the feature extraction stage designed to convert the continuous speech signal into discrete frame features, while capturing distinctive characteristics specific to each speaker. Next, we provide an overview of DNN models utilized in speaker modeling. Finally, we explore the various scoring techniques employed in the ASpR task, accompanied by a discussion of the evaluation tools and metrics used to report system decisions.

2.2 A look back at statistical models

Historically, diverse statistical models played a central role in the evolution of speaker recognition systems. To start with, Vector quantization (VQ), introduced to speaker recognition in the 1980s [17, 18], is a technique that models a speaker using a set of prototype vectors. This technique is firstly used for data compression applications for the purpose of computational speed-up techniques [19]. VQ aims to map a feature vectors space to a set of clusters in that space, each represents a speaker characteristic. The centroids of the clusters are therefore considered as a compressed representation of the feature vectors, namely a codebook, as show in Figure 2.1. This codebook is trained for each speaker using clustering algorithms (e.g. K-means). During inference, a matching score is calculated between a new feature vector and a speaker codebook.

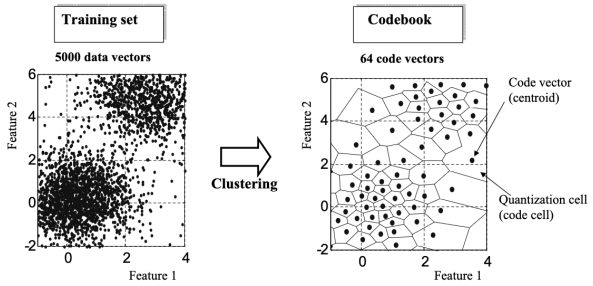


Figure 2.1: The VQ codebook training[20]

In the early 2000s, the Gaussian mixture model-based universal background model

(GMM-UBM) was introduced [15], serving as a foundational framework for speaker recognition for over a decade. The GMM, is a generative speaker model trained in an unsupervised manner. It is constructed of a limited mixture of multivariate Gaussian components. These components represent various spectral features essential for creating a comprehensive speaker model, resulting in a speaker-dependent probability density function (PDF). With GMM, we passed from a discrete prototype vectors modeling in VQ to a continuous representation. Compared to VQ, the probabilistic nature of GMMs has demonstrated their superiority as speaker models, because it permits for a better modeling of variability.

GMM approach was beneficial for speaker identification tasks. For speaker recognition, a reference model is needed for comparison with the claimed speaker's model to make the final decision. This reference model promoted the development of a universal model, also referred to as the universal background model (UBM), firstly introduced in [21]. UBM is a very large GMM trained to capture the speaker-independent distribution of speech features for a broad range of speakers. Building upon it, GMM-UBM approach was proposed [15], to adapt a speaker's GMM model by updating the parameters of the UBM model. GMM-UBM significantly improved performance and opened the door for various representative models, such as support vector machines [22] and joint factor analysis [23].

As statistical models evolved, the notion of GMM *Supervectors* emerged, representing fixed-dimensional vectors for modeling variable-duration utterances. These high-dimensional vectors were typically generated by concatenating the parameters of a GMM model. The field of speaker recognition was then transformed with the proposal of the *identity vector* or i-vector in [16]. I-vectors effectively reduced the dimensionality of these Supervectors into more compact representations. The GMM-UBM/i-vector approach [24] remained the state-of-the-art for speaker recognition for several years, until the emergence of DNN-based ASpR models.

2.3 DNN-based ASpR framework

Using DNN models, the framework of ASpR is mainly composed of two operational phases as illustrated in Figure 2.2. During the training phase, acoustic features are firstly extracted from speech utterances belonging to a predefined set of speakers. These features are fed into a DNN model, trained for a supervised speaker classification task. The goal of this task is to encode and inject more speaker information into the DNN model. The trained model becomes therefore able to generate speaker representations, known as *speaker embeddings*.

In the testing phase, the classifier component is removed from the process. For a speaker recognition task, given two speech utterances, the same process is applied to extract the corresponding speaker embeddings for the two utterances using the trained DNN model. Scoring is then performed by comparing these two embeddings, and the result is compared to a predefined threshold in order to determine the final decision.

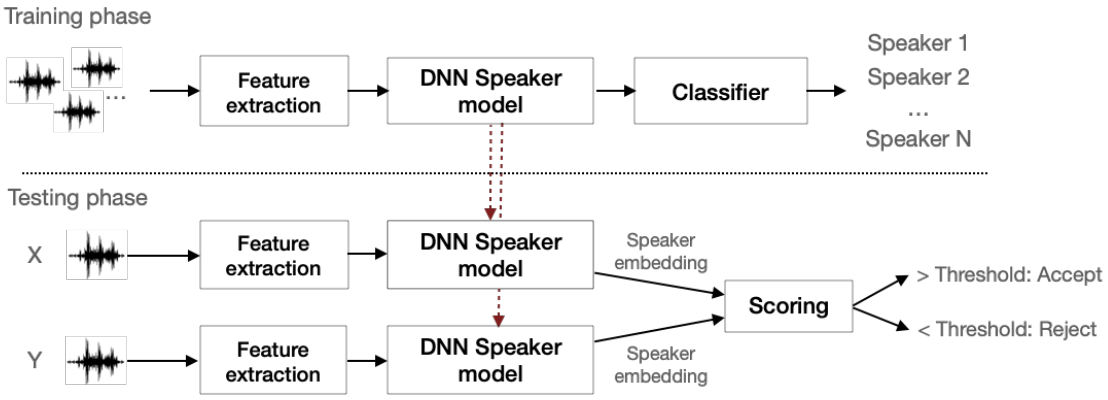


Figure 2.2: DNN-based ASpR framework

In the remainder of this section, we delve into a detailed description of each component within the DNN-based ASpR framework, as illustrated in Figure 2.2. We start by outlining feature extraction methods, categorizing them into two main groups: 1) Conventional handcrafted features, and 2) Self-supervised features extracted using pretrained models. Subsequently, we explore the DNN speaker models, ranging from the classic d-vector model to the most recent advancements. Finally, we present the mostly employed scoring techniques within the DNN-based ASpR framework.

2.3.1 Feature extraction: hand-crafted features

Before we dive into the feature extraction process of handcrafted features, we firstly describe the Voice Activity Detection (VAD) technique followed by other preprocessing steps that precede features extraction. Subsequently, we show particular emphasis on the most commonly used features, namely Filter-banks and MFCC. This focus is justified by our adoption of these features in this work. The whole process is illustrated by Figure 2.3, and further described in the following.

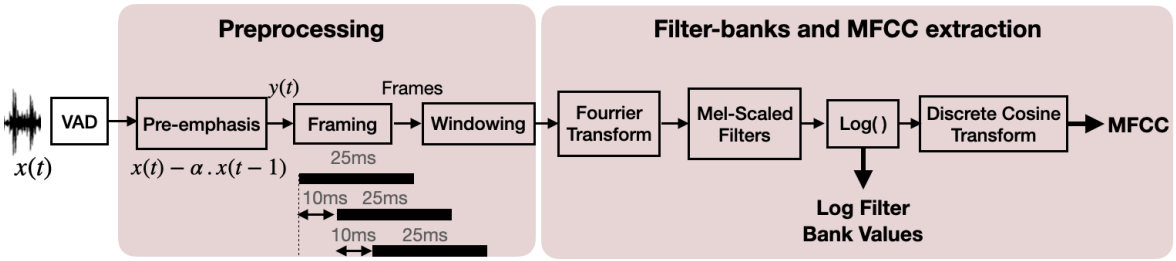


Figure 2.3: Preprocessing and feature extraction block diagram

Voice Activity Detection

It is important to note that VAD serves as a fundamental preprocessing step not only in DNN-based ASpR but also in general speaker recognition systems. VAD is a technology

used in speech processing to distinguish between speech and non-speech segments in an audio signal [25]. VAD algorithms typically consider factors such as energy levels, spectral features, and amplitude variations over time. When speech is detected, VAD marks the corresponding segments as active or "voice," and when non-speech or silence is detected, it marks those segments as inactive [1]. In ASpR systems, VAD helps reduce the amount of irrelevant information, making it easier for the system to focus on the meaningful speech content. The most simple and widely used technique is detection based on energy, namely *Energy-based* technique.

Preprocessing

A preliminary preprocessing stage is essential to prepare the speech signal [26] before feature extraction. This stage primarily involves the conversion of the continuous time-domain speech signal into discrete frames. It is composed of three steps as shown in Figure 2.3. The whole process is described as follows:

- *Pre-emphasis*: In voiced sections of speech, the energy decreases as the frequency increases. Pre-emphasis counteracts this effect by increasing the energy in those segments by employing a high-pass filter with a coefficient denoted as α.
- *Framing*: The speech signal is dissociated into short segments, *Frames*, usually of 25 milliseconds (ms). The overlap between every two consecutive frames is generally 10 ms.
- *Windowing*: Each frame is multiplied by a smooth window function to minimize the signal discontinuities at the beginning and end boundaries. The most popular window function is Hamming.

Filter-bank and MFCC extraction

Various types of features that describe the short term spectral content, have been proposed in the literature [20]. Linear Prediction Coefficients (LPC) [27], Perceptual Linear Prediction (PLP) coefficients [28] and Mel-Frequency spaced Cepstral Coefficients (MFCC) [29] were widely used and are shown to be effective for speaker recognition systems. MFCC is the most commonly used feature extraction method. The extraction steps of these features shown in Figure 2.3 are described as follows [26]:

- *Fast Fourier Transform (FFT)*: The FFT is applied to transform the speech signals from the time domain into the frequency domain. This transformation yields the magnitude frequency response of each frame, which represents how the energy is distributed across different frequencies.
- *Mel-scaled filters*: The N magnitude coefficients are converted to a fewer K Filter-bank outputs. These filters reduce the detailed spectral information including noise, and retain only efficient representation. Typically, triangular filters are employed for this purpose, specifically with a Mel scale. The Mel-scale

aims to replicate the non-linear perception of sound by the human ear, prioritizing discrimination at lower frequencies and reducing discrimination at higher frequencies.

- *Log Filter-bank values*: The logarithm operation is applied to the Filter-bank values. This operation serves two main purposes. First, it expands the scale of the coefficients. Second, it decomposes multiplicative components into additive ones. The outcome of this step is a set of Filter-bank energies, with each channel representing energy within a different frequency band.

- *Discrete cosine transform (DCT)*: This step converts the log Filter-bank spectral values into cepstral coefficients using DCT. The purpose of this step is to decorrelate the Filter-banks coefficients and provide a compact representation of the Filter-bank resulting in the MFCC features.

One additional step to the resulting features is general feature normalization. In the log-spectral (i.e. Filter-banks) and cepstral domains (i.e. MFCC), features are prone to variations due to channel noise that becomes additive. By subtracting the mean vector, feature sets obtained from different channels become zero-mean and the effect of the channel is substantially reduced [30]. This feature normalization technique is the simplest and the mostly used among many proposed techniques in the literature [31, 1].

This resulting spectral or cepstral sequence representation is the starting point for almost all speech-related tasks. In the recent approaches, simple filter-bank energy features are shown to be more effective than MFCC when large neural networks are used for modeling [32]. They are used as an input representation to many DNN approaches, including large pretrained models, such as Whisper [33].

2.3.2 Feature extraction: self-supervised features

With the recent growth in computational resources and capabilities, there has been notable advancement in the development of deep learning feature extraction models, especially in the domain of speech recognition. These features have surpassed conventional handcrafted features by generating highly abstract embedding features directly from audio waveform. These extracted features provide a rich and powerful representation for various subsequent tasks, including speaker recognition. One important advantage of these models is that they learn patterns from large amount of unlabeled audio data in a *self supervised learning* (SSL). The most popular models are Wav2vec and WavLM that we describe as follows.

Wav2Vec

Wav2vec is a SSL model proposed by Facebook AI in 2019 [34] for automatic speech recognition task. The goal was to train a model on huge amount of data without the

need of transcriptions. This model was then trained by learning the difference between original speech examples and modified examples. Recently, in [35], the same company proposed a second version of this model, Wav2vec 2.0, with more complex architecture that includes up to 317 million parameters. The main idea was to mask some parts of the speech input and then try to predict them. This process allows to capture many aspects of the speech signal including speaker traits, noise, etc.

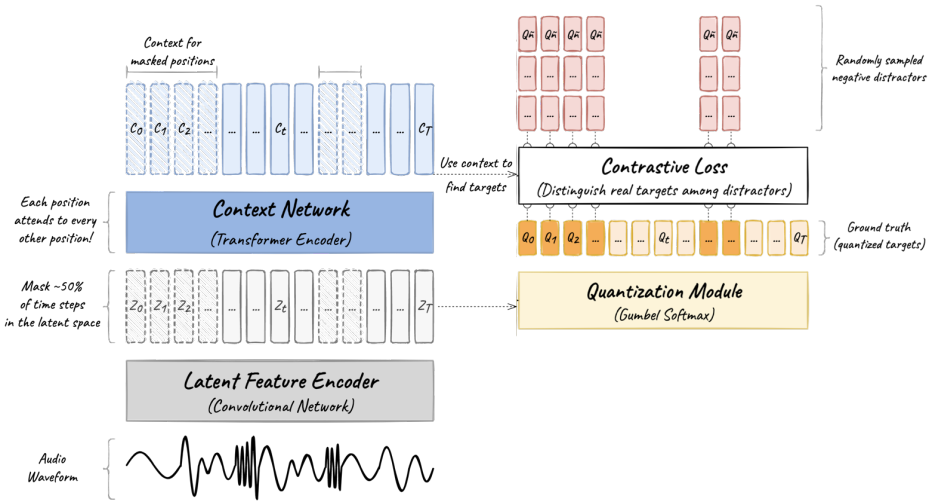


Figure 2.4: The architecture of wav2vec 2.0 model[1]

The model architecture presented in Figure 2.4 is composed of three main modules: the latent feature encoder, the quantization module and the context network. Given a raw waveform of the speech audio X, for each 25 ms, a multilayer convolutional neural network generates latent audio representations Z of 512 dimensions. These representations are then discretized into speech units learned in the quantization module [36]. The transformer encoder [37] takes the latent feature vector with approximately half of the audio representations being masked. Finally, the output of the transformer is used to solve a contrastive task. This task pushes the model to predict the correct discretized speech units for the masked parts of the speech. Table 2.1 illustrates the number of parameters of Wav2vec 2.0 based on the variant [34]. Wav2vec 2.0 Base model comprises around 95 million parameters, while the large model uses ∼316.62 million parameters.

WavLM

WavLM, introduced recently in [38], represents a SSL model designed to acquire a comprehensive speech representation that encapsulates various speech characteristics. The fundamental concept underlying WavLM involves masking noisy or overlapped segments within speech data, after which the model tries to predict the original speech, effectively performing both denoising and prediction tasks. Similar to Wav2Vec, WavLM

[1]https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html

adopts a transformer-based architecture as shown in Figure 2.5. Similarly to Wav2vec 2.0, the scale of WavLM varies, as shown in Table 2.1, with WavLM Base and WavLM Base+ models comprising 94.70 million parameters, while the larger WavLM Large model features ∼316 million parameters [38].

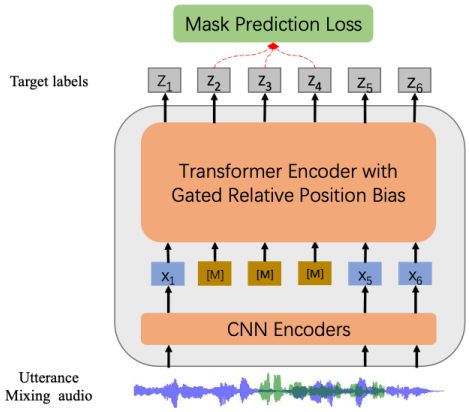


Figure 2.5: The architecture of WavLM model [38]

The remarkable success achieved by WavLM and Wav2Vec models for feature extraction is mainly driven by their ability to handle unannotated data, the complexity of their architectures and the huge number of parameters they incorporate. As is often the case, Table 2.1 shows that larger models tend to exhibit increased complexity and improved accuracy [39]. Notably, as shown in Table 2.1, WavLM outperforms the performance of Wav2Vec 2.0 in speaker recognition task on SUPERB benchmark [40].

Table 2.1: Comparison of speaker recognition performance in terms of EER (defined later in this chapter) between pretrained Wav2vec and WavLM [38] on SUPERB Benchmark [40]

Model	# of parameters	EER (lowest is best)
WavLM Base	94.7M	4.69%
WavLM Base+	94.7M	4.07%
WavLM Large	316.6M	3.77%
Wav2vec 2.0 Base	95.04M	6.02%
Wav2vec 2.0 Large	317.38M	5.65%

2.3.3 DNN speaker model

In the domain of speaker recognition, there has been a dedicated effort to employ DNN models for the direct modeling of speaker characteristics. Typically, these models use spectral or cepstral features extracted from audio waveform as their input (as described in §2.3.1), and they are specifically trained for a speaker classification task (Figure 2.2). The main idea behind these models is to generate fixed-length speaker embeddings for the variable length speech utterances of the speaker. For this purpose,

they adopt a common framework with distinct network architectures, consisting of three key components as illustrated in Figure 2.6. 1) The DNN frame-level extractor which extracts temporal representations from input features. 2) A pooling layer that leverages the information from all the frames to obtain an utterance-level representation of fixed-length, namely the speaker embeddings, 3) A speaker classifier that takes these embeddings as input and use them to classify between different speaker classes during the training phase.

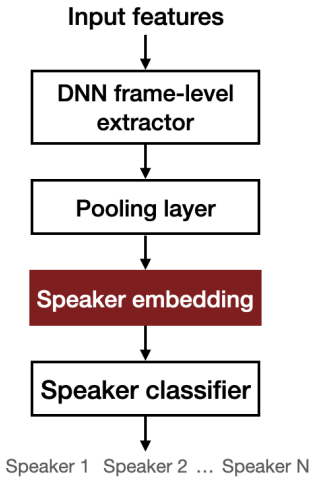


Figure 2.6: The components of a DNN speaker model

In the following, we provide an overview of the architectural designs and models proposed in the literature for each component of the speaker recognition framework.

DNN-based extractor architectures

- *D-vectors*: One of the earliest DNN-based speaker embeddings are, *d-vectors*, proposed in 2014 by Google [41] for a text-dependent speaker recognition task. The model is mainly a feed-forward neural network of multi layers that inputs Filter-banks features of training frames that are stacked together with their surrounding context frames. This network differs from the multi-layer perceptron (MLP) in that it uses a maxout DNN [42] which is a strategy that consists in dropping out some neurons from the layers of the network for optimisation reasons. The speaker representation, denoted d-vector, is therefore obtained by accumulating the activations of the last layer for each frame.

- *Time Delay Neural networks (TDNN)*: Instead of stacking frames at the input of the network, a TDNN architecture is introduced in [43] to handle short-term temporal context. TDNN is a kind of one dimensional convolutional neural network. Compared to DNN-models, the architecture of TDNN is designed to handle the context of the input cepstral acoustic features of frames (i.e. MFCC). It was firstly used in speech recognition in [44]. As described in Figure 2.7 (with gray lines), at each layer of the network, a value is computed using a window of 5

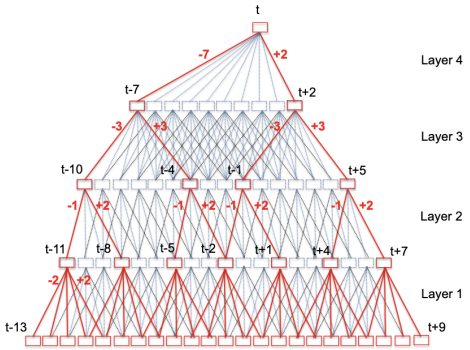


Figure 2.7: TDNN computation diagram [43]

frames, then 4 frames, then 7 frames and then 10 frames respectively for layers 1, 2 , 3 and 4. The network starts by considering a local context at the bottom layers and as long as we achieve upper layers of the network, the considered context become larger. However, as can be noticed, due to the huge amount of computations during training, a subsampling strategy was applied to the layers of the network [43]. This strategy consists in discarding some connections between the units of two successive layers in order to both reduce computation and consider larger context from precedent layers. This subsampling strategy is clearly shown in Figure 2.7 with the red lines across the layers of the TDNN architecture. This architecture allows to capture local and long term temporal correlation between speech frames. The number of parameters involved in the training of the TDNN model are 7.7 million parameters.

TDNN architecture becomes one of the most popular structure for speaker recognition [43, 45]. It is then adopted by the well known x-vector [46, 47], with the configuration shown in Figure 2.8. X-vector system takes as input Filter-banks of 24 dimensions with a frame-length of 25ms, mean-normalized over a sliding window of up to 3 seconds. Given a speech utterance of T frames, the first five layers consider increasingly temporal context for frame number t. Subsequently, a statistical pooling layer leverage information from all T frames and calculates a mean and a standard deviation. The output of this layer represents the speaker embedding, namely the x-vector.

Layer	Layer context	Total context
frame1	$[t-2, t+2]$	5
frame2	$\{t-2, t, t+2\}$	9
frame3	$\{t-3, t, t+3\}$	15
frame4	$\{t\}$	15
frame5	$\{t\}$	15
stats pooling	$[0, T)$	T
segment6	$\{0\}$	T
segment7	$\{0\}$	T
softmax	$\{0\}$	T

Figure 2.8: TDNN architecture configuration for x-vectors [47]

Motivated by the success of x-vectors, many variants based on an improved TDNN architecture are later developed. For instance, an extended TDNN architecture (E-TDNN) was introduced in [48], which outperforms the x-vector baseline [47]. It is trained with a slightly wider temporal context than TDNN [49, 50]. The authors in [51] proposed factorized TDNN (F-TDNN) to reduce the number of parameters for training. It aims to factorize the weight matrix of each TDNN layer into the product of two low-rank matrices. It also constrains the first low-rank matrix to be semi-orthogonal to prevent loss of information when reducing dimensions [49, 50]. Some components were combined with TDNN architecture for better results. For example, [52] added a statistic pooling after each layer of the TDNN architecture to manage the variation of temporal context in the frame-level transformation. Many works combined TDNN with other DNN models to enhance performance and capture more information at different levels such as TDNN-LSTM [53], TDNN-BLSTM [54], CNN-LSTM-TDNN [55].

- *Residual networks (ResNet)* [56]: The main difference between ResNet architecture and a standard multi-layer CNN is the skip connections or *Identity shortcut connections* added to the CNN blocks, as shown in Figure 2.9. It allows the model to skip one or more layers. The main goal of these skip connections is to address the problem of gradient vanishing due to very deep neural network. A residual block is composed of two 2-dimensional CNN layers separated by a Rectified Linear Unit (ReLU) activations. The input of the residual block is added to its output in order to constitute the input of the next residual block.

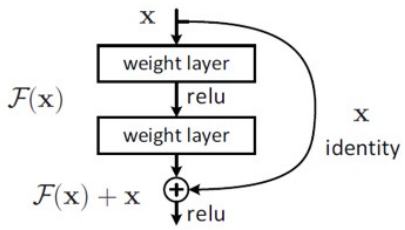


Figure 2.9: Identity Shortcut connection [56]

Figure 2.10 describes the architecture of a variant of ResNet, ResNet34, used to extract x-vectors [57]. The input of this ResNet is 40 Filter-banks features. Filter-bank features are very commonly employed as input for ResNet by most of the works [57, 58, 59, 60, 61]. The first step before entering the common layer behavior is Conv2D-1, consisting of a convolution, batch normalization and max pooling operation. Then the 4 residual blocks all follow the same strategy. They perform 3x3 convolution with a fixed feature map dimension [32, 64, 128, 256] respectively, bypassing the input every 2 convolutions. Each convolutional layer is followed by a batch normalization layer and a ReLU activation function.

Many works investigated over ResNet in speaker recognition field such as [62, 63, 64, 65, 66]. Some other works modified over the baseline ResNet architecture of x-vectors [60, 61, 65, 66, 67] focusing both on dealing the dependency between frames as well as the interdependence of the channels (i.e. the feature dimension).

For instance, [68] introduced a block to the ResNet named *Squeeze and excitation* (SE) block. The idea behind this block is to learn the interdependence between the channels, and give weights to some channels in order to highlight informative features and remove less useful ones [69, 70].

Layer name	Structure	Output
Input	–	$40 \times 200 \times 1$
Conv2D-1	3×3, Stride 1	$40 \times 200 \times 32$
ResNetBlock-1	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$, Stride 1	$40 \times 200 \times 32$
ResNetBlock-2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 4$, Stride 2	$20 \times 100 \times 64$
ResNetBlock-3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 6$, Stride 2	$10 \times 50 \times 128$
ResNetBlock-4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$, Stride 2	$5 \times 25 \times 256$
StatsPooling	–	10×256
Flatten	–	2560
Dense1	–	256
Dense2 (Softmax)	–	N
Total	–	–

Figure 2.10: ResNet configuration for x-vectors [57]

- *ECAPA-TDNN*: It is a variant of TDNN architecture that incorporates skip connection property of ResNet and the SE blocks to enhance the performance of the x-vector. It is firstly proposed in [71]. It aims to produce finer-grained features extracted at multiple scales of the network. Figure 2.11 illustrates the structure of SE-Res2block, being the main component in the ECAPA-TDNN architecture.

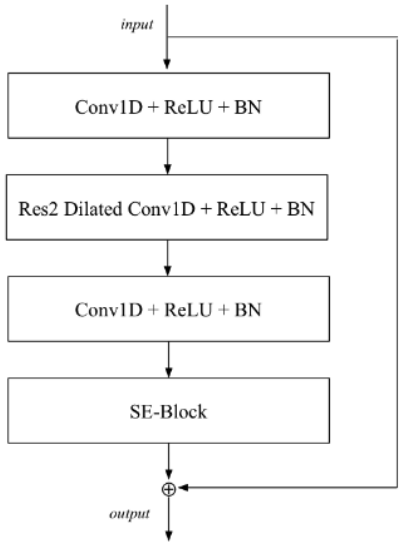


Figure 2.11: Structure of the SE-Res2Block of the ECAPA-TDNN architecture [71]

For each frame, two dense layers are enveloping a dilated convolutional layer that

is meant to gradually constructing the temporal context. The first dense layer serves for feature dimensionality reduction while the second one restores the original dimension of features. This is followed by a SE-block to model interdependency between channels. As the number of channels increases, the performance of ECAPA-TDNN are shown to be better [71, 72]. Each SE-Res2block involves a residual connection. As shown in Figure 2.12, features from each SE-Res2block are then aggregated in a Multi-layer Feature Aggregation (MFA) fashion.

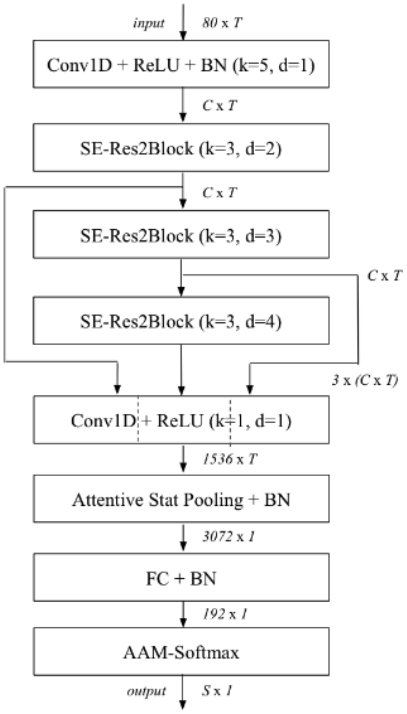


Figure 2.12: The architecture of ECAPA-TDNN [71]

- *MFA-Conformer*: It is proposed by [73] to handle local and global context features. Its architecture is based on a combination of Transformer and CNN. Given that the transformer could effectively capture long term context, the CNN is shown to be very good at capturing details about features and local context.

Table 2.2 illustrates a comparison of speaker recognition performance between the different architectures of speaker model on VoxCeleb1-O dataset [57]. WavLM large model presents the best performance on VoxCeleb1-O compared to other models. MFA conformer is shown to outperform ResNet and ECAPA-TDNN. Additionally, the use of Wav2vec 2.0 features as input to TDNN [74] is proved beneficial for speaker recognition performance.

Table 2.2: Comparison of speaker recognition performance in terms of EER between different model architectures on VoxCeleb1-O

Model	# of parameters	EER (lowest is best)
TDNN	7.7M	1.46%[43]
ResNet34	23.2M	1.03%[73]
ECAPA-TDNN	20.8M	0.82%[73],1.01%[38]
MFA-Conformer	20.5M	0.64%[73]
Wav2vec 2.0-TDNN(XLS-R 1B[75])	265M	0.69%[74]
WavLM Base+	94.7M	0.84%[38]
WavLM Large	316.6M	0.617%[38]

Pooling layers

The pooling layer serves as an intermediate between frame-level layers and utterance-level layers of the DNN-based speaker models. It aims to aggregate, in some way, all variable length temporal information to produce fixed-length representation, namely *speaker embedding*. In the following, we present the most used pooling layers in DNN-based speaker model.

- *Statistical pooling*: This is the most typical and classical pooling method that appeared with the x-vectors architecture[47, 57]. It computes the mean and the standard deviation of the frame-level representations, concatenates them and propagates them through the segment-level layers of the network. The work in [76] used different pooling layers such as statistical measures (maximum, mean, standard deviation, skewness, kurtosis...) to evaluate ResNet performance for different classification tasks. The authors in [77] found that the use of standard deviation in pooling layer improves the results, highlighting that dynamic information encoded by standard deviation not only contains the phonetic information but also provides speaker-dependent information.

- *Attentive pooling*: In the frame-level representations, [78, 79] showed that only some set of features are more involving in discriminating speakers than others. Thus, many works incorporated attention mechanism in the pooling layer to give more importance to some frames with respect to others. It computes attentive weights for each frame, while focusing on the most important frames to discriminate speakers. [80] introduced an attentive statistics pooling method, which computes importance-weighted standard deviations and weighted means of frame-level features using attention mechanism. Following the same mechanism, other works such as [62, 81, 82] mainly focused to improve the quality of the aggregation of the pooling layer for better performance. The attention mechanism drives and orients temporal information in the speaker classification direction which adds more interpretability to the speaker embedding regarding temporal information.

Classifier loss functions

DNN-based speaker recognition systems usually adopt classification-based objective functions to classify between classes of speakers. The utterance-level representation extracted from pooling layer is propagated through a fully connected layer for the classification step.

- *Softmax*: The Softmax loss is typically used for deep embedding. For instance, in x-vectors[46] and d-vectors [41] extractors, the objective function used is the minimum cross entropy which takes the Softmax as the output layer.
- *Angular Softmax*: The Softmax loss is very effective in maximizing the between-class distance, but lacks constraints to minimize the within-class variance. A new variant of Softmax was proposed in the field of face recognition [83, 84], namely *angular softmax* (ASoftmax). It introduces an angular margin between embeddings of different classes [85]. In a study conducted by Xiang et al. [86], the importance of three different Angular-based losses in producing distinctive speaker embeddings was investigated. ASoftmax losses offer several advantages over the softmax loss [87]; they transform the learned feature distribution in an angular fashion, which is well suited with similarity scoring methods like cosine similarity during speaker recognition inference. Furthermore, these losses help minimize within-class variance by introducing an angular component to have better control over the decision boundaries between speaker classes. Consequently, ASoftmax loss has become the state-of-the-art choice for many DNN-based models dedicated for speaker recognition task.

2.3.4 Scoring

The scoring component evaluates the similarity between two speaker embeddings stemming from two speech utterances, and then compares this score with a threshold. Linear Discriminant Analysis (LDA), Probabilistic Linear Discriminant Analysis (PLDA), and cosine distance are the widely used scoring techniques for speaker recognition.

- *Linear Discriminant Analysis (LDA)*: LDA is a supervised dimensionality reduction and classification technique. It aims to find a linear combination of features that is more effective in discriminating between classes. It projects the data to a lower dimensional subspace in a way that it maximizes the variability between classes and minimizes the variability within the classes [88]. The covariance matrices of between (S_b) and within classes (S_w) are illustrated by Equation (2.1) and Equation (2.2) respectively.

$$S_b = \frac{1}{S}\sum_{s=1}^{S}(\mu_s - \mu)(\mu_s - \mu)^T \tag{2.1}$$

$$S_w = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(x_{i,s} - \mu_s)(x_{i,s} - \mu_s)^T \tag{2.2}$$

Where S is the total number of speakers, μ_s is the mean of samples at the speaker s, μ is the mean of all samples, $x_{i,s}$ is an utterance belonging to a speaker s and n_s if the number of utterances of speaker s. The projections of LDA are found by optimizing Equation (2.3) [89]:

$$V = \frac{A^T S_b A}{A^T S_w A} \tag{2.3}$$

Where A is a projection matrix.

- *Probabilistic Linear Discriminant Analysis (PLDA)*: PLDA is a probabilistic variant of LDA that handles unseen classes [90]. PLDA assumes that speech samples are following a Gaussian distribution[91]. A classical Gaussian PLDA model implies that a speaker embedding is constructed as follows:

$$x = m + V_y + z \tag{2.4}$$

Where m is the mean of the speaker embeddings, y is the speaker latent variable, and z is normally distributed with zero mean and full covariance matrix. To estimate V_y and the covariance of z, PLDA uses Expectation maximization algorithm [92].

For a speaker recognition task, the verification score between two speaker embeddings is therefore calculated using the log-likelihood ratio of two hypotheses. H_0 assuming that both embeddings are coming from the same speaker, and H_1 assuming that they are coming from different speakers. This score is modeled as follows:

$$\text{score} = \log \frac{p(x_1, x_2|H_0)}{p(x_1, x_2|H_1)} \tag{2.5}$$

Where x_1 and x_2 are two speaker embeddings. If the score is greater than the threshold, then both embeddings come from the same speaker, otherwise, the speakers are different. The PLDA assumes a general format of speaker embeddings as in Equation. (2.4).

From Equation (2.5), and Equation (2.4), the PLDA score is calculated as following:

$$\begin{aligned} PLDA_{score} = log\mathcal{N}(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} m \\ m \end{bmatrix} \begin{bmatrix} \Sigma_{tot} & \Sigma_{ac} \\ \Sigma_{ac} & \Sigma_{tot} \end{bmatrix}) \\ -log\mathcal{N}(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} m \\ m \end{bmatrix} \begin{bmatrix} \Sigma_{tot} & 0 \\ 0 & \Sigma_{tot} \end{bmatrix}) \end{aligned} \tag{2.6}$$

where $\Sigma_{tot} = VV^T + \Sigma$, $\Sigma_{ac} = VV^T$ and m is the average of all speaker embeddings.

- *Cosine distance*: The cosine distance is mainly computed from the cosine similarity. Cosine similarity determines the angle between two vectors in a high-dimensional space. In [93] the cosine similarity measure-based scoring was pro-

posed for speaker recognition. Given a comparison pair, the cosine score compresses two speaker embeddings into one single value. Cosine similarity is inversely proportional with cosine distance and it can be written as:

$$cos(\theta) = \frac{A.B}{||A||.||B||} \tag{2.7}$$

In contrast to PLDA, cosine similarity scores do not necessitate the training of a back-end model. With DNN models trained with a ASoftmax objective function, it has been found that Cosine similarity outperforms PLDA scoring for speaker recognition (Table 2.3) which is not the case for i-vectors [93]. This is promoted by the use of ASoftmax. [94] found that the PLDA is more robust to test data with different acoustic conditions and domain mismatch as illustrated in Table 2.3.

Table 2.3: Comparison of scoring backends of ResNet model using EER for two test datasets

Scoring	**Dataset**	**EER (lowest is best)**
Cosine	VoxCeleb1	1.06% [94]
PLDA	VoxCeleb1	1.86% [94]
Cosine	CNCeleb1(domain mismatch)	10.11% [94]
PLDA	CNCeleb1(domain mismatch)	8.90% [94]

2.4 Evaluation protocols and metrics

The speaker recognition system is evaluated based on the accuracy of the speaker model. This accuracy can be determined from the false acceptance ratio (FAR) and false rejection ratio (FRR). FAR (Equation (2.8)) represents the proportion of times the system incorrectly accepts an input as a match (i.e. target). Conversely, FRR (Equation (2.9)) is the percentage of times the system erroneously rejects an input as a non-match (i.e. non-target or imposter) when it is an actual match trial. In a speaker recognition system, the process involves comparing two speaker embeddings stemming from two speech samples by calculating a similarity score. To reach a conclusive decision, the system employs a predefined threshold. This threshold serves as a reference point for comparing the scores and determining whether the comparison pair corresponds to a target or an imposter.

$$\text{False Acceptance Rate (FAR)} = \frac{\text{Number of FA errors}}{\text{Number of non-target trials}} \tag{2.8}$$

$$\text{False Rejection Rate (FRR)} = \frac{\text{Number of FR errors}}{\text{Number of target trials}} \tag{2.9}$$

The Equal Error Rate (EER), the Detection Cost Function (DCF) and the Detection Error Trade-off (DET) curve are commonly used evaluation metrics in the speaker

recognition[2] literature. A description of these evaluation metrics is provided in the remaining of this section.

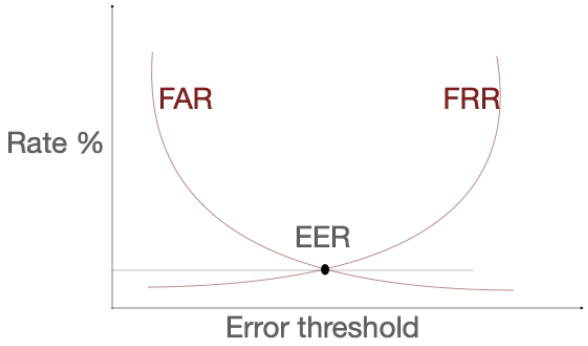


Figure 2.13: Relationship between FAR, FRR and EER

2.4.1 Equal Error Rate

The EER determines a posteriori the threshold value where false acceptance rate and its false rejection rate are equal. The point at which the rates are intersected, as shown in Figure 2.13, is referred to as the equal error rate (EER). The lower the EER value, the better the recognition system.

2.4.2 Detection Cost Function

The Detection Cost Function (DCF) was introduced by NIST, where each type of error can be penalized differently [95]. In scenarios like banking authentication systems, where security is of paramount importance, a system that exhibits a bias towards rejecting a target user might be preferable than the one that easily accepts users. As a result the DCF is considered as application dependent measure. The DCF is computed as the sum of weighted sum of FRR and FAR for a given threshold τ as follows:

$$DCF(\tau) = C_{FR}P_{FR}(\tau)P_{target} + C_{FA}P_{FA}(\tau)(1 - P_{target}) \tag{2.10}$$

Where C_{FR} and C_{FA} are the penalties of FR and FA errors respectively. P_{target} is the prior probability of target speaker. P_{FR} is the probability of FR given that the pair is target and the threshold is τ and P_{FA} is the probability of FA given that the pair in non target and a threshold τ.

[2]The Log Likelihood Ratio Cost (C_llr) is also a commonly used metric and it is defined in the next chapter.

2.4.3 DET curve

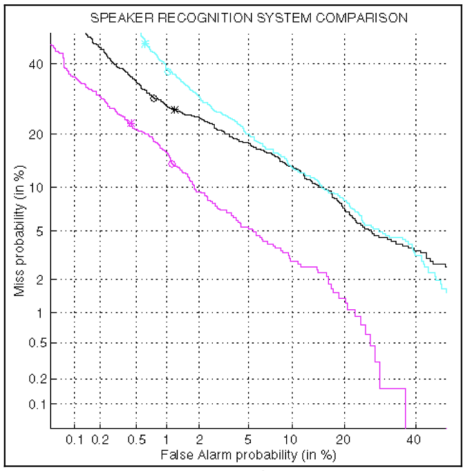


Figure 2.14: Plot of DET Curves for a speaker recognition evaluation [96]

The DET curve, as introduced in [96], serves as a graphical tool for illustrating the trade-off between FAR (i.e. False alarm probability) and FRR (i.e. Miss probability), as illustrated by Figure 2.14. What sets it apart is its use of a uniform scale for both axes, which results in a spread-out plot and enhanced differentiation among systems with varying performance. As a system's performance improves, the curve gradually moves closer to the origin. Similarly to DCF, the FAR can be minimized by increasing the detection threshold to a significant level, this comes at the cost of a higher FRR. This is generally very dependent on the application of the system [95].

2.5 Summary

In this chapter, we provided an overview of various state-of-the-art approaches used in each component of the speaker recognition framework. This comprehensive summary allows us to establish the foundational choices underpinning our work for each stage of the ASpR framework.

In the context of the feature extraction stage, we highlighted that traditional handcrafted features offer a user-friendly and computationally efficient alternative when compared to deep speech representations such as WavLM and Wav2vec. Additionally, these handcrafted features are more readily interpretable and explicit, in contrast to deep representations, which encode a vast amount of speech-related information that remains abstract and largely unexplored. In the realm of speaker modeling stage, we detailed the progression of DNN-based methods designed to extract speaker embeddings. For the sake of performance, these models have been embracing progressively complex architectures and increasing number of parameters to effectively address the diverse forms of speech variability and deliver better results. Nevertheless, they often sacrifice the transparency of information flow and the comprehensibility of their

architecture.

In line with the state-of-the-art DNN speaker recognition systems, this thesis adopts Filter-bank features in the initial version of embeddings extraction phase. As long as the main focus of this thesis is providing informing and interpretable ASpR system, the initially proposed solution of our work is based upon the baseline ResNet architecture with a standard deviation pooling and ASoftmax objective. From a critical perspective, the choice of ResNet is may be not the most accurate, but it is thought to be a good compromise between performance and complexity (Table 2.2). Also, we agree that employing of an attentive pooling instead of a statistic pooling could be more advantageous in terms of interpretability. Introducing attention at this stage would enhance interpretability and facilitate the localization of important frames for the training task. This remains a consideration for future perspectives.

Chapter 3

Forensic application of automatic speaker recognition

3.1 Forensic Automatic Speaker Recognition 30
3.2 Bayes paradigm assessing the value of evidence 31
3.2.1 From frequentist to Bayesian approach 31
3.2.2 The Bayesian interpretation and the court 32
3.3 Centrality of likelihood ratio 33
3.3.1 LR interpretation 34
3.3.2 LR estimation from similarity scores 34
3.3.3 Calibration of LR into well-calibrated LR 36
3.4 On the use of automatic methods in forensic science 36
3.4.1 Acceptability by the court 37
3.4.2 A help or a burden for forensic scientist? 38
3.4.3 Requirement for explanations in forensic science 39
3.5 Summary 40

In the previous chapter, we provided an overview of the fundamentals of ASpR systems, highlighting the remarkable progress of DNN-based models to achieve greater performance. In the present chapter, our focus shifts to the practical application of ASpR in the specific field of forensics. We concentrate on adapting ASpR systems to the Bayesian framework to assess the value of speech evidence in the judicial process. Subsequently, we highlight the central role of the Likelihood Ratio and its interpretation by the court. Lastly, we point out the challenges associated with the use of ASpR models in forensic applications.

3.1 Forensic Automatic Speaker Recognition

Forensic science, is a multidisciplinary field that involves the application of various scientific and investigative techniques to gather, analyze, and interpret physical evidence found in a crime scene. The role of forensics is the provision of accurate and reliable information that can be used in legal proceedings. *Forensic Speaker Recognition (FSpR)*, also referred to as *Forensic Voice Comparison*, is the procedure of determining whether a specific individual, namely *the suspect speaker*, can be identified as the source of a provided voice recording, known as *the trace*. In this practice, a forensic expert examines the two recordings, conducting a comparative analysis [97]. Then, the value of this evidence is reported in the form of a Likelihood Ratio (LR).

FSpR addresses the challenges associated with forensic speech material, which introduces additional complexities to the general variability of speech. These challenges include short voice records, low voice quality, background noise, and uncontrolled forensic conditions such as screaming over the phone or a speaker disguising their voice [98] (refer to Figure.3.1). Additionally, FSpR aligns with the specific requirements for presenting the value of evidence in a courtroom setting. Specifically, it requires the adaptation of the general speaker recognition process to the Bayesian framework, often seen as the "logically and legally correct" framework [99] for the court's interpretation of the weight or the value of evidence.

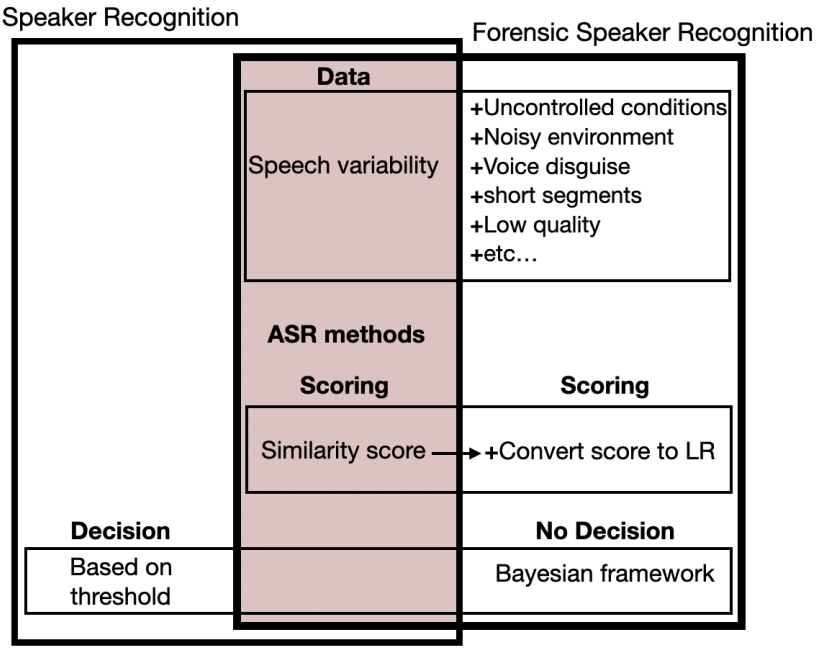


Figure 3.1: Speaker recognition Vs. Forensic Speaker Recognition

Forensic Automatic Speaker Recognition (FASpR) refers to the use of ASpR methods by forensic practitioners in the process of evaluating the value of evidence [100]. ASpR models aim to compare two recordings and return a score. This score is used in ASpR systems to make decisions about whether the two recordings belong to the

same speaker. In a forensic context, this score is used to estimate the LR under the Bayesian framework (Figure.3.1), assessing the value of evidence. It is essential to ensure that the final decision is ultimately deferred to the court only. FASpR field relies and usually follows the progress of ASpR methods developed in non-forensic context. From the use of GMM-UBM models [101, 102] to the use of x-vectors [103] until the very recent use of ECAPA-TDNN [71] in a real forensic voice comparison case [104, 105]. Despite the limited literature on the use of DNN models in FSpR [106, 107, 108], some software solutions based on DNN-based approaches have been developed and adopted by legal institutions. For instance, VOCALISE, a commercial product of Oxford Wave Research introduced in 2012 [109], has incorporated x-vectors [47] in its latest version released in 2019 [103]. Notably, this biometric software has received support and collaboration from institutions such as the German Bundeskriminalamt, the Netherlands Forensic Institute (NFI), and the UK Ministry of Defence. Another commercial product, Phonexia Voice Inspector [110], uses DNN models to provide police forces and forensic experts with a speaker recognition tool. This system has been used and evaluated by the German Federal Criminal Police[1]. However, while automatic methods are increasingly integrated into the judicial process, its use within the forensic context should not be taken for granted and requires a high level of caution.

In the next sections, we further detail the forensic field, presenting the frequentist paradigm and the shift toward the Bayesian paradigm actually employed by FSpR systems. Subsequently, we draw attention to the centrality of the LR, including the approaches for its estimation and its interpretation in the legal context. Finally, we highlight the potential risks associated with the use of automatic and AI-based approaches in forensics, emphasizing the need for cautions.

3.2 Bayes paradigm assessing the value of evidence

Two common strategies for assessing the evidence in a forensic context are the frequentist and Bayesian approaches. In the specific forensic context, such approaches serve to evaluate the value of evidence from crime scene and report the conclusions to the court. In this section, we mainly describe both approaches, while highlighting the paradigm shift from frequentist to Bayesian framework followed by the Bayesian interpretation of the evidence by the court.

3.2.1 From frequentist to Bayesian approach

According to [111], the key distinction between Bayesian and frequentist approaches lies in their focus: the Bayesian approach considers the probability of hypotheses, whereas the frequentist approach solely deals with the probability of observed data.

In the frequentist approach, a single hypothesis is examined, positing that the sample belongs to a specific dataset. It is then compared to a "null hypothesis" that

[1]https://ondatashop.com/phonexia-voice-inspector/

assumes the data is a result of random chance. The probability of this event occurring by chance is calculated. If this probability is exceptionally low (below a predetermined threshold), the null hypothesis is rejected. This decision supports the hypothesis under investigation [101]. The frequentist approach has faced criticism from proponents of the Bayesian methodology, as discussed in [111, 112, 99]. One fundamental challenge with the frequentist approach is that forensic experts must make assumptions about prior probabilities, which are often based on convention instead of considering the specific circumstances of the case. An "uninformative prior" is frequently employed, assuming that each possible explanation is equally likely. Furthermore, the choice of the alternative hypothesis, which represents the odds against a match occurring by chance, may not always be appropriate for a given case.

In contrast, in the Bayesian approach, when interpreting evidence, it is important to consider the context [113] and not just focus on one hypothesis. Instead of solely looking at the probability of one scenario (i.e the questioned sample coming from the suspect), the expert must also think about the probability of the evidence in the context of alternative scenarios (i.e. the questioned sample coming from someone else). This helps in evaluating how strongly each scenario is supported by the evidence. This approach is now the commonly used among all forensic disciplines, becoming the standard framework [99].

3.2.2 The Bayesian interpretation and the court

In a FASpR scenario, the court is faced with a decision-making under uncertainty [114]. The judge needs to know how much likely the differences or similarities between two speech samples (i.e. the evidence) prove that the suspect speaker has or has not produced the trace sample [97, 114]. This is determined by the ratio of the conditional probability at Equation (3.1).

$$\frac{P(H_p|E)}{P(H_d|E)} \tag{3.1}$$

- H_p represents the prosecution hypothesis which states that the two speech samples under comparison are coming from the same source speaker.
- H_d is the defence hypothesis which states that the two speech samples belong to different speakers.
- E is the evidence from the crime scene like DNA trace, a vocal recording...

The solution of this statement is given by the Bayes theorem in Equation (3.2) which says that the posterior odds are determined from a combination of the prior odds (i.e. the use case related information) and a new data which is the questioned sample. Evett et.al [113] states that this equation is *". . .the fundamental formula of forensic science interpretation"*.

$$\underbrace{\frac{P(H_p|E)}{P(H_d|E)}}_{\text{Posterior odds (Role of the court)}} = \underbrace{\frac{P(E|H_p)}{P(E|H_d)}}_{\text{Likelihood ratio (Role of the expert)}} \cdot \underbrace{\frac{P(H_p)}{P(H_d)}}_{\text{Prior odds (Given by the court)}} \tag{3.2}$$

The prior odds in Equation.3.2 represent the view on the prosecution (H_p) and the defence (H_d) hypotheses before seeing the evidence. The posterior odds can be seen as an update of the prior odds in light of knowledge of the scientific evidence. This update is done by multiplying the prior odds by the LR corresponding to the evidence. The LR is the ratio of the likelihood that the trace and suspect speech samples have the same source, and the likelihood that they come from different sources.

The LR estimate is the responsibility of the forensic expert and it summarizes his statement (Equation.3.2). Then, it is up to the court to evaluate its worth and decide whether to take it as an aid to their decision or not [114]. The court is responsible for estimating the prior odds based on the case related information and then determines the posterior odds using the LR provided by the expert. The expert is not able to make an estimation of the probability of a hypothesis such as these two samples were/were not spoken by the same speaker using only the evidence, which is stressed out by many researchers in the field. For instance, Aitken in [115] states:

> *"It is very tempting when assessing evidence to try to determine a value for the probability of guilt of a suspect, or the value for the odds in favour of guilt and perhaps even reach a decision regarding the suspect's guilt. However, this is the role of the jury and/or judge. It is not the role of the forensic scientist or statistical expert witness to give an opinion on this. . . . It is permissible for the scientist to say that the evidence is 1000 times more likely, say, if the suspect is guilty than if he is innocent."*

This is also supported by Champod and Meuwly in [99] that *"the analysis of the scientific evidence does not allow the scientist alone to make an inference on the identity of the speaker"*. It is therefore the role of the court to make its decisions based on the strength of evidence reported by the forensic expert. Rose also agreed in [116] that *"It is neither logically nor legally correct for the forensic expert to attempt to state the probability of a hypothesis given the evidence"*.

3.3 Centrality of likelihood ratio

Within the forensic science community, the LR is commonly adopted as the *"logically and legally correct"* framework to evaluate and present the strength of evidence to the court [99, 117, 118]. In this section, we focus on the interpretation of the LR followed by the commonly employed methods for its computation, primarily based on similarity scores, derived from ASpR models. Lastly, we describe the evaluation of performance of these LR methods.

3.3.1 LR interpretation

The LR is the statement of the support degree for the prosecution hypothesis against the defence hypothesis [119]. Referring to Equation(3.2), the numerator of the LR, $P(E|H_p)$, quantifies the degree of similarity between the trace and the suspect samples, while the denominator, $P(E|H_d)$, quantifies the degree of typicality of both the trace and suspect samples in the relevant population. As the similarity between two samples increases, the likelihood of them originating from the same speaker also increases, resulting in a higher ratio. However, this must be counterbalanced by their typicality. The more typical the samples, the more probable it is that they were randomly selected from the relevant population, leading to a lower ratio [114]. The value of the LR is, therefore, a result of the interplay between these two factors: similarity and typicality [99].

The LR value is the degree of support of one hypothesis versus its alternative. In simpler terms, if the $LR > 1$ then the evidence supports the prosecution hypothesis, but if the $LR < 1$ the evidence supports the defence hypothesis. Furthermore, assuming a $LR = 3$, this means that the evidence supports the prosecution hypothesis H_p three times more than the defence hypothesis H_d. Thus, a single LR value is self sufficient, in contrary to a similarity score from ASpR system, that may have a meaning only if it is compared to a predefined threshold or to another set of scores. It is important to note that the LR supports a belief about the hypotheses, but it is in no case a belief about the hypotheses [120, 119]. For this reason, with the LR only, one could not make any decision.

3.3.2 LR estimation from similarity scores

Across different branches of forensic, the estimation of LR could be performed mainly by two methods: score-based approach [121, 122, 123] or feature-based approach [121, 123, 122, 124, 125]. Feature-based approach, also referred to as the direct method [100, 101], calculates the LR as the ratio of two density functions of the feature vectors of the comparison pair directly under prosecution H_p and defence H_d hypotheses. An example of the use of this approach in FASpR is the traditional i-vector modeling followed by PLDA scoring described in §2.3.4 [100]. Score-based approach estimates the LR as the ratio of the two likelihoods of the scores under H_p and H_d hypotheses.

In FASpR systems, the most commonly used approach is score-based, because of its ease of implementation and its robustness face to feature variations. It is composed of two main steps: 1) a similarity score is calculated between two representations, the trace sample and the suspect sample, extracted using any ASpR system, 2) the score is transformed into a likelihood ratio [120]. This approach could be easily plugged into any DNN-based ASpR model that outputs a score. The transformation of scores to LR values could be done by three main methods described and further detailed as follows.

- *Probability density estimation*: LR computation in forensics has been classically performed modeling the hypotheses-conditional distribution of the scores (i.e.

univariate distributions under each hypothesis) [101]. It is described mathematically as follows:

$$LR = \frac{P(S|H_p)}{P(S|H_d)} \tag{3.3}$$

The probability density function $P(S|H_p)$ is the intra-variability distribution. Its evaluation gives a measure of the probability density of observing the evidence under H_p. The $P(S|H_d)$ in the denominator is the inter-variability distribution, and its evaluation gives a measure of the probability density of observing the evidence under H_d [119]. The generally used is a Gaussian distribution [126]. This is the less recommended method for score transformation as explored in [119].

- *Pool Adjacent Violators (PAV)*: Firstly introduced in [127], it transforms scores into a set of LR values. PAV algorithm could be trained on a set of training scores under prosecution and defence respectively, then apply the trained transformation on the score of a new data. This requires ground truth labels (i.e. target/non-target) of the training scores [119].

- *Logistic regression*: Firstly used for LR computation in[127, 128]. It aims to obtain an affine transformation by shifting and scaling a set of scores in order to optimize an objective function. Given a set of score S as evidence, this affine transformation can be defined as follows:

$$f_{lr} = \log(\frac{P(H_p|S)}{P(H_d|S)}) = a + b \cdot s = \log(O(H_p|S)) \tag{3.4}$$

Where $O(H_p|S)$ is the posterior odds that could be defined in function of H_p only, since H_p and H_d are complementary events. It is expressed as follows as:

$$O(H_p|S) = \frac{P(H_p|S)}{1 - P(H_p|S)} \tag{3.5}$$

The Bayes'theorem gives the logarithm of the LR for a given prior $O(H_p)$:

$$\log(LR) = a + b \cdot s - \log(O(H_p)) \tag{3.6}$$

This leads us to the logistic model:

$$P(H_p|S) = \frac{1}{1 + e^{-f_{lr}}} = \frac{1}{1 + e^{-\log(LR) - \log(O(H_p))}} \tag{3.7}$$

The coefficients (a, b) are learned during model training. For target scores we may define the obtained value as $f_{lr}^{t} = a + b.s^{t}$ and for non-target scores $f_{lr}^{nt} = a + b.s^{nt}$. Then, the coefficients are learned by making $P(H_p|S)$ as close as possible to 1 for target trials and to 0 for non-target trials [120].

The logistic regression has been found to be more robust to overfitting and dataset shift than PAV in forensically realistic conditions [119].

3.3.3 Calibration of LR into well-calibrated LR

Transforming similarity scores into LRs is not enough to say that we obtained well calibrated LRs. Additionally, the obtained LRs should be also as best calibrated as possible. *Calibration* is a property of a set of LRs, where the LR is interpreted as a measure of the weight of evidence [119, 129]. A LR system is well calibrated when the output gives calibrated LR values [130, 131]. In contrary to the probability density estimation that often gives non calibrated LR values, PAV transformation and logistic regression methods transform the scores into directly calibrated LRs [119]. Ill-calibrated LRs necessitates a further post-hoc step of calibration that yield a better calibrated LRs [130]. This calibration could be also performed by either logistic regression or PAV algorithm.

In forensic science, there are some specific performance metrics that are stated to be adequate and commonly used to evaluate LR computation and calibration. The Log Likelihood Ratio Cost ($Cllr$) has been proposed in speaker recognition [127] and then applied in forensic [132, 133] to measure the performance of LR values. The lower is the value of C_{llr} the better is the performance. This value is defined as follows:

$$C_{llr} = \frac{1}{2.N_p} \sum_{i_p} \log_2(1 + \frac{1}{LR_{i_p}}) + \frac{1}{2.N_d} \sum_{j_d} \log_2(1 + \frac{1}{LR_{j_d}}) \tag{3.8}$$

Where i_p and j_d are summing over the N_p LR values given H_p is true and over N_d LR values given H_d is true, respectively. LR_{i_p} and LR_{j_d} are the likelihood ratios derived from test pairs known to be target and non-target comparisons respectively. N_p and N_d are the number of target and non-target comparisons respectively. In [127], PAV algorithm is used to decompose C_{llr} as follows:

$$C_{llr} = C_{llrmin} + C_{llrcal} \tag{3.9}$$

Where C_{llrmin} represents the discrimination cost of the LR method, whereas C_{llrcal} is the calibration cost of the system. The C_{llr} is referred to also as C_{llract} which is the actual C_{llr}.

3.4 On the use of automatic methods in forensic science

The use of AI models for FASpR systems is an emerging field progressing at a modest rate. For instance, in the Dutch criminal justice system, FASpR has been in use at the Netherlands Forensic Institute (NFI) since December 2018 only [134]. Before this date, no general decision about its admissibility was required. Now, it is used in almost a third of the forensic speaker comparison cases undertaken by the NFI [134]. Police agencies also have already embraced FASpR systems that are unregulated and operate as black boxes, in collaboration with technology companies. However, this raised a lot of ethical concerns about its trustworthiness in such high-risk field [135, 136, 137].

In this section, we draw attention to the acceptability of all judicial parties the use of automatic systems in the forensic context. Firstly, we recapitulate the different

positions taken by the courts regarding the general use of automatic methods in criminal cases. We then emphasize the responsibility of forensic scientist when employing automatic models to assess the strength of evidence. Lastly, we highlight the need for caution and the requirement for interpretable and explainable AI models within the critical field of forensic science.

3.4.1 Acceptability by the court

To illustrate the positions taken by the courts regarding the use of automatic AI models in the criminal justice system [137], we propose a categorization into three distinct positions. This proposed categorization is presented in Figure 3.2. 1) **Over reliance**: Certain judges blindly accept the value of evidence, even when aware that it is derived from an automatic system. 2) **No trust**: Others neglect the value of evidence when assessed using a black box method. 3) **Reasonable reliance**: Some other courts express concerns about the use of black box systems in criminal cases, demanding more comprehensive explanations regarding the assessment of evidence.

One example of the first position is a court in Pennsylvania[2] that is responsible for reviewing a defense challenge. The court denied the defense's request for an external, independent evaluation of the software used in the case. This is further detailed in §.A.1.1. Another example to add here is some other courts that have also admitted a specific software, asserting its reliability without providing a clear rationale for permitting its use[3]. The courts have assumed it sufficient that the software developer claimed to have validated the software. Further details are in §.A.1.1.

A one telling example of the second position is a ruling in 2019[4], described in §.A.1.2, where a state trial judge determined that it was a mistake to rely on such forensic evidence. The judge further proposed that any convictions stemming from the use of this software should undergo a review. The judge underscored that the software was essentially a "black box," since no independent expert was allowed to examine its internal workings.

In an example showing the third position, a federal judge made an unusual decision to compel the Office of the Chief Medical Examiner in New York City to reveal the source code of its probabilistic genotyping software, which was employed for analyzing DNA mixtures [138]. This action led to the emergence of various concerns regarding the software's accuracy, ultimately resulting in its discontinuation [138]. Another example[5], described in §.A.1.3, in the same direction mentioned in [139] where the judge considered that *"without access to value-added equations, computer source codes, decision rules, and assumptions, teachers could not exercise their constitutionally-protected rights to due process"*.

[2]Commonwealth v. Foley, 38 A.3d 882 (Pa. Super. Ct. 2012)

[3]U.S. v. Russell, No. CR- 14- 2563 MCA, 2018 WL 7286831, at *8 (D.N.M. Jan. 10, 2018).

[4]People v. Thompson, 65 Misc. 3d 1206, 2019 N.Y. Slip Op. 51521, 118 N.Y.S.3d 383 (N.Y. Sup. Ct. 2019)

[5]Local 2415 v. Houston Independent School District, 251 F. Supp. 3d 1168 (S.D. Tex. 2017), p. 17.

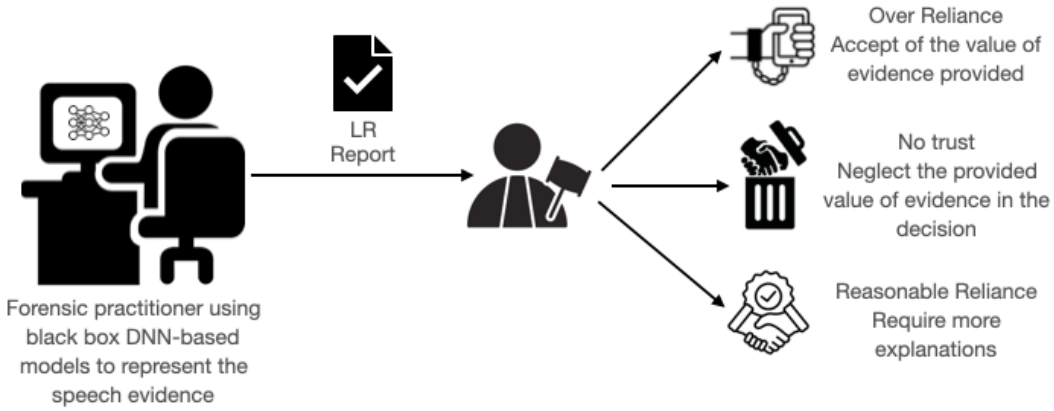


Figure 3.2: Our proposed categorization of the different positions of the court regarding the use of automatic methods

Even if there is no common applied regulations about the use of DNN models to assess the value of evidence, this does not mean that these models are not prone to error or to discrimination bias. Judges should not only assess the reliability of the expert's chosen method but also the consistency of its application to the specific facts at hand [139, 135, 136].

3.4.2 A help or a burden for forensic scientist?

Is the use of AI models in evaluating the evidence a help or a burden for forensic scientists? Does it entail greater responsibility or offer an escape from it?

Enchanted by the magic of DNN-based models, forensic practitioners, particularly software engineers, increasingly rely on the high performance of these models. However, having a DNN model that has demonstrated remarkable accuracy on certain datasets is insufficient to justify trust in its reliability, nor to assume it will perform equally well on different types of data. What if the model has acquired a bias towards a particular accent or nationality? Given that the forensic practitioner had no intention of introducing such bias, and often may not even be aware of it, who should be held responsible for this?

The absence of awareness among forensic practitioners regarding the potential for bias is unacceptable in such a critical field [137, 140]. The role of a forensic practitioner extends beyond merely assessing the strength of evidence and presenting conclusions in court. He should also be able to provide explanations for the value assigned to evidence and the process used to derive it [141]. This includes details such as the training data used for the model, the patterns it has learned, the information encoded in its generated representations, and whether it exhibits biases toward specific classes, among other considerations. The authors in [137] formulated the required explanation that the forensic practitioner should provide through three questions: *What were the main factors in a decision? Would changing a certain factor have changed the decision? Why did two similar-looking cases get different decisions, or vice versa?*. It is not only

the judge and jury who have the right to inquire about the trustworthiness of the AI model in use, "the prosecution and defense also have a fundamental right: the right to explanation" [142, 137]. That is why, the forensic scientist should always be able to provide explanations about any stage of the evaluation process when requested.

3.4.3 Requirement for explanations in forensic science

The integration of explainable AI (XAI) in forensics have only recently begun to attract attention. Some works were published from the 2020's to the present to emphasize the significance of AI interpretability in forensic context and criminal justice [143, 142, 135, 136, 144, 145, 139, 146, 147, 140, 141]. For instance, some very recent works [142, 139] focus on the issues of black box AI models and the importance of XAI models used in digital forensics. The work in [144] clarifies the legal requirement of explainability in judicial decisions. The authors emphasized this requirement by the European Court of Human Rights *"In accordance with Article 6 (refer to §A.2.1) of the Convention, judgments of courts and tribunals should adequately state the reasons on which they are based"*. This is also justified by some European countries having such obligations in their regulations such as in Belgium[6] (refer to §A.2.2). The authors also mentioned the example of the United Kingdom, where it is a common law principle that a judgment must be reasoned, meaning that it *"explains to the parties and to any wider readership why the judge has reached the decision he has made."* If a judgment is not sufficiently explained, it can be vacated by a higher court. As a result, the authors in [144] proposed four required levels of AI explainability in forensics such as providing the main features used in an output, providing all the processed features, providing a comprehensive explanation of the output and providing an understandable representation of the whole model. In [148], the author explored the necessity and methodology behind providing explanations for a specific judicial decision, taking into account various factors such as the decision's significance and the decision-maker's role. The author also emphasized that when a judge issues a criminal sentence, which is one of the most crucial decisions in court, it is imperative to provide an explanation to enable the defendant to identify and address any potential impropriety or error. [137] focused on the significance of regulating AI models to ensure accountability in a forensic context.

The demand for interpretability and explainability[7] of black box models within the forensic context remains an open issue, lacking clear implementation in practical rules and regulations. Consequently, existing literature is only restricted to addressing the necessity and prerequisites without delving into the practical application of explainability methods on currently employed automatic techniques.

[6] art. 149 of the Constitution of the Code on judicial proceedings

[7] These two terms are discussed and defined in Chapter 4

3.5 Summary

In conclusion, within this chapter, we outlined the use of ASpR methods in the field of forensic science, addressing the related challenges. We specifically introduced the Bayesian framework considered as the logically and legally accepted approach for presenting the value of evidence to the court. Within this framework, we demonstrated the adaptation of ASpR methods output, transforming similarity scores into a straight interpretable LR value by the court.

Moreover, we draw a particular attention to a noteworthy challenge, concerning the use of DNN black box models for assessing the value of evidence in FASpR system. This challenge involves the risk of discrimination bias in the forensic context. We pointed out that all stakeholders in the judicial process, including the court and forensic practitioners, should be aware of this risk. Cautions should be exercised when employing AI and DNN-based models and it is imperative that all parties take responsibility for the system's output, offering comprehensive explanations for the model's findings and decisions.

Chapter 4

Interpretability and explainability in AI

4.1 Introduction . . . 42
4.2 Context of explainable AI . . . 42
4.2.1 Awareness of explainability . . . 43
4.2.2 The need of explainability . . . 43
4.3 Terminology . . . 45
4.3.1 Interpretability . . . 45
4.3.2 Explainability . . . 46
4.3.3 Other related concepts . . . 47
4.4 Literature review: Taxonomy . . . 47
4.4.1 Local Vs. Global . . . 48
4.4.2 Inherently Vs. Intrinsic Vs. Post-hoc . . . 48
4.4.3 Model-specific Vs. Model-Agnostic . . . 49
4.5 XAI for speech system decision . . . 53
4.6 Challenges . . . 54
4.7 Conclusion . . . 54

In the precedent chapters, we firstly presented an overview of state-of-the-art ASpR systems based on DNN models, highlighting their inherent complexity, that often leads to opaqueness. Then, we described the particular use of DNN-based ASpR systems in a critical field such as forensics and we emphasized the need for cautions and the requirement for explanations when using black box models. In this chapter, our focus shifts towards the adoption of AI interpretability and explainability methods, aiming to address the opaqueness in DNN models, within the specific forensic context and in a broader sense.

4.1 Introduction

Artificial intelligence (AI) is incorporated in many applications of our daily lives. It is becoming an indispensable actor that drive our decisions. AI is involved in the most critical domains and life-changing decisions such as disease diagnosis and criminal accusation. However, this shift towards AI-driven decision-making has arisen a lot of critical inquiries regarding trust, transparency, and ethical issues. Though AI models are shown to be increasingly powerful, they are not without limitations. The most significant one is the opaqueness or the lack of transparency of highly complex architectures. This opaqueness prevents human from being able to verify or understand the reasoning of the system and how particular decisions are made. Systems whose decisions cannot be well-interpreted, should not be easily trusted, especially in fields, such as healthcare or forensics, where moral and fairness issues are of paramount concern.

The imperative to enhance the transparency and interpretability of AI models has given rise to the field of Explainable AI (XAI). In the following sections, we start by defining the context of XAI, emphasizing the growing awareness and the demand for interpretable systems. Then, we define a terminology of the related concepts as defined in the literature. This is followed by the taxonomy of XAI methods in the broader domain of AI. Afterwards, we specifically review the existing literature on XAI in speech tasks. Finally, we address the challenges and limitations faced in the field.

4.2 Context of explainable AI

The core objective of XAI [149] field is to develop techniques and methods that makes AI models more comprehensible to humans, while preserving their predictive performance. The term "Explainability" has its origins dating back to the late 1980s, when there was an exploration of the ability of expert systems to provide explanations for their decisions [150]. Nevertheless, with the emergence of powerful AI models in the last decade, the focus was mainly oriented towards performance enhancement, while neglecting the ability of these models to explain their decisions. With the rapid development of AI models in critical domains where the stakes are extraordinarily high, XAI field regained attention. Figure 4.1, generated by Google Trends[1], shows the evolution of the research interest of the two terms "interpretability" and "explainability" from 2004 until now. This growing interest is also illustrated by many calls from international organizations to make AI explainable. In the following, we illustrate this awareness by some examples and we highlight the need for XAI in some particular fields.

[1] https://trends.google.fr/trends/

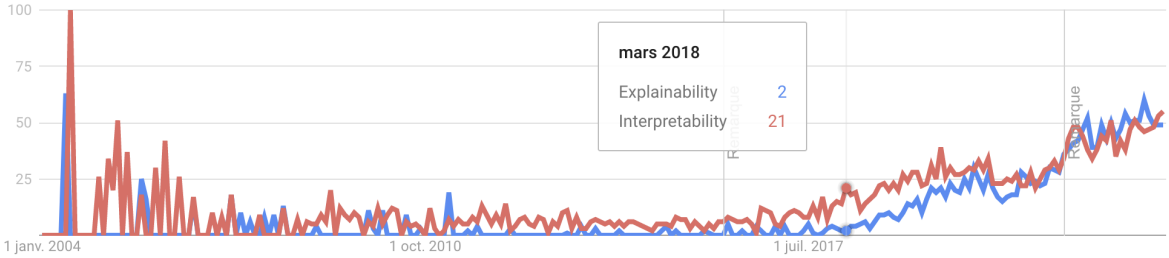


Figure 4.1: Google trends for research interest of the terms "interpretability" and "explainability" from 2004 until September 2023

4.2.1 Awareness of explainability

The prevalence of AI models in various aspects of our daily lives has fueled an increasing demand for ensuring transparency and interpretability within these models. As a response, many initiatives and policy changes have emerged. In 2016, the White House Office of Science and Technology Policy released the U.S. report on AI titled "Preparing for the Future of Artificial Intelligence", focusing on AI systems that are open, transparent, and comprehensible to facilitate decision justifications for individuals [151]. In 2018, the Pentagon allocated $2 billion to its "AI Next" initiative, led by the Defense Advanced Research Projects Agency of the United States Department of Defense (DARPA), aiming to promote AI research and development. This initiative marked the beginning of the XAI field [152].

Moreover, the European General Data Protection Regulation (GDPR), announced in 2018, specifically in articles §.A.3.1 and §.A.3.2, that data subjects are provided with "The right to be informed". They have the right to an explanation of algorithmic decisions, including the provision of a list of contributing factors to the decision upon request [153, 154, 155]. Following the GDPR regulations, the European parliament announced in 2021, the European Union's Artificial Intelligence Act (AI Act) [156]. This regulatory framework aims to ensure an ethical and responsible AI that respects the fundamental rights and societal values. It also regularizes the use of AI in high-risk systems such as Law enforcement and encourages to promote transparency in these systems. This is pronounced by Recital 38 in §.A.4.

4.2.2 The need of explainability

One of the most important motivation behind providing explainability is the adoption of AI models in many high-stake (i.e. high-risk) applications. in these particular contexts, the risk of encountering a wrong prediction is substantial. In the following, we show that providing interpretability and explainability is an ethical imperative in three most critical fields.

- *Healthcare sector*: Medical diagnosis models carry the responsibility of human lives. Inaccurate diagnoses resulting from wrong decision factors can have harmful

consequences, and subsequently a misguided patient treatment. The need for interpretability in these models is evident. It ensures that the rationale behind medical diagnoses is transparent for the medical practitioner and comprehensible for patients who have the right to understand the reasoning behind their diagnoses and treatment plans. A lot of works in the literature show interest in applying XAI methods in healthcare applications [157, 158, 159].

- *Banking sector*: In banking applications such as voice-based authentication, granting the access to a wrong user is very dangerous. Customers need to have confidence in the security and reliability of voice authentication, especially when it comes to accessing their financial accounts. For instance, a report of BBC in 2017 detailed how a UK bank (HSBC) was forced to suspend its speaker recognition technology upon uncovering that a malicious actor could deceive the system by employing a vocal recording of the account holder [160]. The system exhibited a notable rate of unauthorized approvals, enabling an unauthorized individual to gain access to the account holder's funds. Thus, this indicates that it is essential to provide explanations of how voice authentication model works and how the decision is made by the system to build trust with customers [161, 162]. Furthermore, regulatory requirements, such as Know Your Customer and Anti-Money Laundering [163] guidelines, necessitate a clear understanding of the decision-making processes behind voice authentication systems. Another famous banking application involves predicting loan approval through machine learning models [164]. According to the regulation B of the *Equal Credit Opportunity Act (ECOA)*[2], all loan applicants have the right for explanations (refer to §A.5). Moreover, this regulation prohibits creditors from discriminating against credit applicants based on factors such as color, religion, sex, etc. This underscores the imperative to provide explanations for the used models to avoid any risk of discrimination bias [165].

- *Forensic science and legal sector*: Another very critical domain of AI application, encountered in the previous chapter, is the legal system. Here the consequences of any misleading decision are highly severe, such as the cases where people are incorrectly denied parole [166] or incorrect bail decisions by the court that allows the release of potentially dangerous criminals [167]. Unfortunately, there is already clear evidence on existing biased systems that produced discriminatory decisions in the criminal justice [8]. In 2016, ProPublica's examination [9] of Correctional Offender Management Profiling for Alternative Sanctions tool, COMPAS, revealed that although the algorithm's overall accuracy is similar for both white and black defendants (i.e. individuals accused with a crime in a legal proceeding) ($\approx 62\%$), the types of errors it makes differ. It tends to categorize black defendants more frequently as high-risk when they are not, while it more often categorizes white defendants as low-risk when they are not. Similarly, another predictive algorithm, PredPol, used by law enforcement, has faced criticism for its potential to disproportionately target low-income, black communities [10]. Such discrimination raises serious ethical questions regarding whether to trust these systems which have decisions over people's lives [139]. An article published

[2]ECOA law signed in U.S 1974 described in appendix A

in The New York Times in 2017 [166] entitled, *"When a Computer Program Keeps You in Jail: How Computers are Harming Criminal Justice"*, illustrates the impact of using AI in criminal justice with some real cases of denied parole people. Transparency of the decision-making produced by an AI-based system in the legal context is therefore a real MUST, even at the cost of accuracy [145, 168]. In the literature, still there is very few works that are dedicated to apply explainable methods on automatic AI-based systems used in the legal context [169, 170].

Another perspective of the need of explainable and interpretable AI models is a technological necessity. Gaining a deep understanding of how AI models function can yield new insights and fresh perspectives on the underlying data. The authors in [171] present a categorization of the need for explainability and interpretability, identifying four key axes from a technological perspective. These axes include the need 1) to explain to justify model decision, 2) to explain to control over the model behaviour, 3) to explain to improve upon existing models and 4) to explain to discover new ideas and new axes. However, explanations are not for free [172]. Generating explanations or making an AI model interpretable takes time and effort and could lead to a decrease in performance. Therefore, the utility of explanations must be balanced between the cost of generating them and the requirement of providing them.

In this thesis, we focus on the domain of forensics and criminal justice, pointed out in the previous chapter, which represents a particularly sensitive field requiring explanations and interpretations for decisions made by automated systems.

4.3 Terminology

So far, the terms "Interpretability" and "Explainability" were used interchangeably. Nevertheless, it is important to note that the definitions of these two terms exhibit important distinctions in the literature of various fields such as social science [173], philosophy [174], psychology [175]... This distinction is subject to a philosophical dilemma where no universally accepted definition adopted of both terms. On the other hand, other researchers do not distinguish between the two concepts and prefer to use them interchangeably [176]. In the following, we report the definitions of both concepts exactly as presented in the literature, followed by some related concepts that we find relevant.

4.3.1 Interpretability

The definitions of interpretability in the literature are very ambiguous and do not have any clear formalism behind [177, 178]. Considering the challenging aspect of this subject, we went for a categorization of these definitions based on our own interpretation. Please note that the following categorization is a simplification for the purpose of clarity and does not provide an exhaustive representation of the diverse viewpoints about this dilemma.

1. **The ability of a model to be understandable**: Fiok et. al in [179] define interpretability as *"the model's ability to present its decisions in terms that humans can understand"*, that is agreed by [180, 181]. Another very recent definition in the same category is the one of [136] *"we refer to predictive models whose calculations are inherently capable of being understood by people"*. Rudin provides a more clear definition in [145] that says *"interpretable ML[3] focuses on designing models that are inherently interpretable"* that is fully adopted by [178].

2. **Using interpretability methods or human reasoning**: This category groups definitions that could define interpretability as, the methods applied to interpret a model in a way that is understandable, or by human reasoning. [182] says that *"The goal of interpretability is to describe the internals of a system in a way that is understandable to humans"*. Also Doshi-Velez and Kim define interpretability of AI systems as *"the ability to explain or to present in understandable terms to a human"*, which is also adopted by [137]. Molnar in his book [183] precises that interpretability *"refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans"*. Adadi and Berrada in [171] agree with Linardatos in [184] that interpretability could be defined as *"the more interpretable a machine learning system is, the easier it is to identify cause-and-effect relationships within the system's inputs and outputs"*. In the same direction, Montavon et. al in [185] say that it is *"the mapping of an abstract concept, for instance, a predicted class, into a domain that the human can make sense of"*.

3. **Human understandability of the model**: From social science perspective, Miller in [173] defines interpretability as *"the degree to which a human can understand the cause of a decision"*. We classify this definition in a different category because we think that it depends on the human ability to understand things and that it is very relative to human judgement even in the case of two humans of the same expertise.

4.3.2 Explainability

Regarding the definition of explainability, this controversial is less important where a group of authors in literature agree that explainability is describing and clarifying the internals of the model. For instance, Linardatos in [184] defines explainability as it *"is associated with the internal logic and mechanics that are inside a machine learning system. The more explainable a model, the deeper the understanding that humans achieve in terms of the internal procedures that take place while the model is training or making decisions"*. Following the same direction, Fiok et. al in [179] define it as a seeking to clarify the model's functioning. Similarly to Rudin [145] and Garret [136] definition in that says that *"explainable ML is a tool permitting to provide post hoc explanations for existing black box models"*.

[3] interpretable machine learning

Differently, Molnar in his book [183] says that *"an explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way"*. Another definition adds another perspective to defining explainability in [181], which is the relationship between explainability and the target of the explanations *"Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand."*.

4.3.3 Other related concepts

Due to the diversity of interpretations in the field, various related concepts have intersected with the interpretability and explainability terms, each offering another perspective derived from both terms. For the sake of precision and in relation with this work, we found it useful to explicitly define and distinguish some of these related concepts, as outlined below:

- *Transparency*: Is the capacity of the model to explain its own functioning even when it behaves unexpectedly [181, 185, 177]. Recently, Garret et.al in [136] discussed this idea by differentiating between *Transparency* and *Interpretability* saying that they are different and that a model could be transparent but not interpretable or interpretable but not transparent.
- *Causality*: The ability of a method to clarify the relationship between input and output in a specified context of use [177, 186, 187].
- *Understandability*: Is the property of a model to make a human understand its operation without elucidating its internal structure or the internal operations by which the model processes data [187, 177].
- *Informativeness*: The ability of the method to provide useful information to the end-user via its output [177, 187].

At this stage, we do not aim to firmly establish our position regarding the terminology of interpretability and explainability. Instead, in the next section, we continue to use both terms interchangeably, and we provide a clear definition towards the end of this chapter.

4.4 Literature review: Taxonomy

In this section, we aim to present the taxonomy of interpretability and explainability methods by providing illustrative examples from the literature of each class. Figure 4.2 presents a taxonomy that we propose based on the existing works [171, 181, 188, 184, 176, 189].

Explainability/interpretability could be either provided under a globally or locally scoop. It could be also divided into post-hoc explanations, intrinsic explanations or

inherently interpretable models. Post-hoc (i.e. after training of the model) explainability could be further classified into model-specific or model agnostic, while intrinsic is by definition a model-specific interpretability. In the following, we define each of these boxes as shown in Figure 4.2.

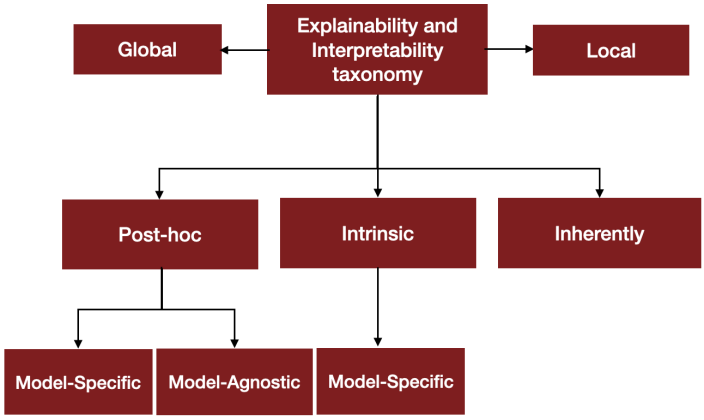


Figure 4.2: Our proposed explainability and Interpretability taxonomy

4.4.1 Local Vs. Global

Local interpretability/explainability consists in explaining the reasons behind a particular prediction of the model [171, 184]. This is used to understand the decision of the model regarding a single instance. Global interpretability/explainability on the other hand provides an overview about the general model behavior given all the predictions [187, 171]. This permits to having an entire view of the model.

4.4.2 Inherently Vs. Intrinsic Vs. Post-hoc

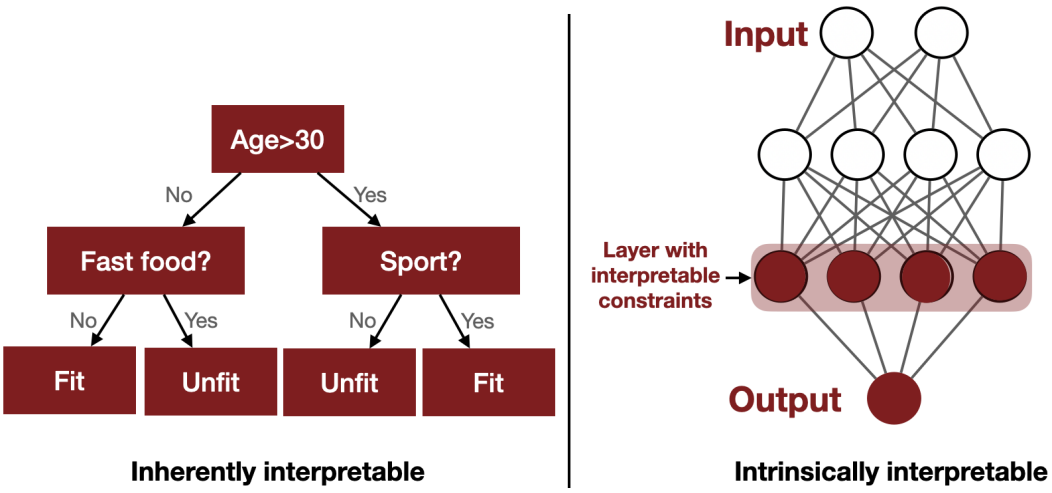


Figure 4.3: Inherently Vs. Intrinsically interpretable/explainable models

Inherently Interpretable models are models that are interpretable by nature. These models exhibit a white box architecture and referred to as *Transparent models* [190, 180, 149]. Logistic regression model, Decision Tree, Rule sets [191] are some examples of this family. These models provide *global interpretability* and they are designed to be simple and basic for understandability purposes. However, they are very limited and do not effectively reflect real-world interactions. Figure 4.3 shows a simple example of a Decision Tree where we can clearly see how we predicted the output (Fit/Unfit) from understandable features like (Age, Sport, and Fast food).

Intrinsic interpretability/explainability on the other hand, is achieved by applying some constraints in the model architecture during training to orient them to be explainable, as shown in Figure 4.3. Such constraints could be by imposing sparsity, monotonicity, causality, or physical constraints that come from the domain knowledge [145, 176]. This type of methods is therefore often *globally* interpretable/explainable and it is very relative and application-dependent. Figure 4.3 shows clearly the distinction between inherently and intrinsically interpretable/explainable.

Post-hoc interpretability/explainability refers to methods that are applied on the model after training to explain its behavior or its decision [176]. It is noteworthy that these methods could be also applied on the two previously mentioned types (i.e. inherently interpretable models or on intrinsically interpretable models) to enhance the interpretability/explainability of the model. Intrinsic interpretability could be achieved, by definition, using model-specific methods, while Post-hoc could be performed, both globally and locally, using either *model-specific* or *model-agnostic* methods as described in the following.

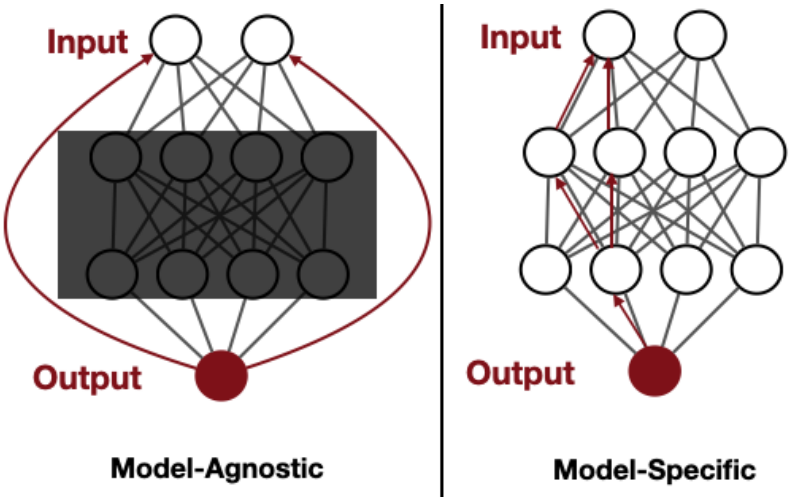


Figure 4.4: Post-hoc explainability methods

4.4.3 Model-specific Vs. Model-Agnostic

Model-specific interpretability is restricted to some particular model architectures, and it could be local and/or global. Among the most known methods in the literature, there

is a collection of methods specific to neural networks. They are mostly based on the back-propagation of the gradient into the neural network architecture to feature the impact of any change of the input with respect to the output such as Integrated-Gradients [192, 193], Guided backpropagation [194], Grad-CAM [195] applied to CNN models. In the same direction, [196] proposed Layer-wise Relevance Propagation (LRP) to be applied on more complex architecture. Starting from the output neurons going back through the lower layers of the network to the input-layer neurons where each neuron redistributes to the lower layer the same amount of information received from the higher layer. All model-specific methods generally provide local interpretations.

On the other hand, model-agnostic methods are designed to be plugged to any model in order to explore its behavior and its output. Among these models, we describe, in the following, the commonly used ones with a particular focus on Shapley values [197] as they are mainly used in this thesis.

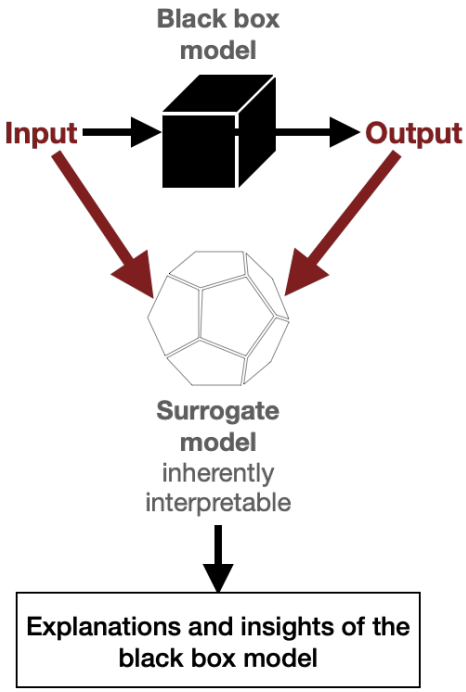


Figure 4.5: Train a surrogate model on input and output of black box model

- *Surrogate models*: The fundamental concept of this method is to create a simplified model that closely mimics the predictions of the complex black-box model while also offering interpretability [183, 176]. There is a common confusion[4] regarding the use of simpler models to interpret black-box models: why create a complex black-box model when we could opt for directly interpretable models like Linear Regression or Decision Tree? The reason is the inherent trade-off between accuracy and interpretability. While interpretable models are straightforward, they often sacrifice accuracy when dealing with intricate data. Complex models, such as deep learning models, excel at handling complexity but lack the inter-

[4]https://maheshwarappa-a.gitbook.io/explainable-ai-1/model-agonistic-methods/surrogate-model

pretability of simpler models. Surrogacy aims to bridge this gap by developing a simple model that can replicate the predictions of the black-box model, balancing accuracy and interpretability. It gives some insights about the black box model functioning. Figure 4.5 describes how a surrogate model could be used for explanations. It takes the input and predictions of the black box models and tries to provide explanations. Surrogate models could be either global or local (§.4.4.1). Local surrogate model is used by a popular explanation method, namely LIME, described in the following.

- *Local Interpretable Model-Agnostic Explanation (LIME)*: Proposed in [198] to provide local explanations. It approximates a black box model with a local surrogate model to explain each single prediction [183]. This method proves particularly effective when the emphasis is on comprehending individual predictions rather than analyzing a set of predictions made by a model.

- *Shapley Additive explanations (SHAP)*: A very powerful tool proposed by Lundberg and Lee in [199]. This approach employs Shapley values from coalitional game theory, as introduced in [200], to equitably allocate gains among players who have contributed unevenly. The game, in this context, represents a prediction task for a specific instance, with the players being the feature values associated with that instance. These players collaborate to share the gains fairly. The *Shapley value*, quantifies the marginal contribution of a feature value to the overall prediction across all conceivable "coalitions" or feature subsets. It determines the share that each "player" (or feature) receives after the game and is mathematically defined as follows:

$$\phi_i(x) = \underbrace{\frac{1}{|F|!} \sum_{S \subset F/\{i\}}}_{\text{Average}} \underbrace{|S|!(|F| - |S| - 1)!}_{\text{Weight}} \underbrace{[f_{S\cup\{i\}} x_{S\cup\{i\}} - f_S x_S]}_{\text{Marginal Contribution}} \tag{4.1}$$

Where x is the input instance, $\phi_i(x)$ is the shapley value for the feature i for the input x for the model f. F is the set of all features while S is a subset of features. f_S is the trained black box model on the subset of features S. $f_{S\cup\{i\}}$ is the trained model on the subset S and feature i. x_S is the x representation using the subset of features S. $x_{S\cup\{i\}}$ is same but using S and feature i. $F/\{i\}$ is the set of all features without feature i. Equation (4.1) shows that the Shapley value of feature i is a weighted average of the marginal contributions of i over all subsets S of F. The first part is averaging over all possible combinations. For a subset S, the weight is the product of the number of permutations of S and the number of permutations of the complement of S and i. The marginal contribution analysis the difference in the output by including or excluding feature i.

Table 4.1: Axioms definition of Shapley values

Properties	**Definition**
Efficiency	The sum of the Shapley values of all features equals the value of the prediction trained with all the features, so that the total prediction is distributed among the features.
Symmetry	The contributions of two feature values should be the same if they contribute equally to all possible coalitions.
Linearity	If two models described by the prediction functions f and g are combined, the distributed prediction should correspond to the contributions derived from f and the contributions derived from g.
Dummy	The features that do not contribute should have a Shapley value of 0.

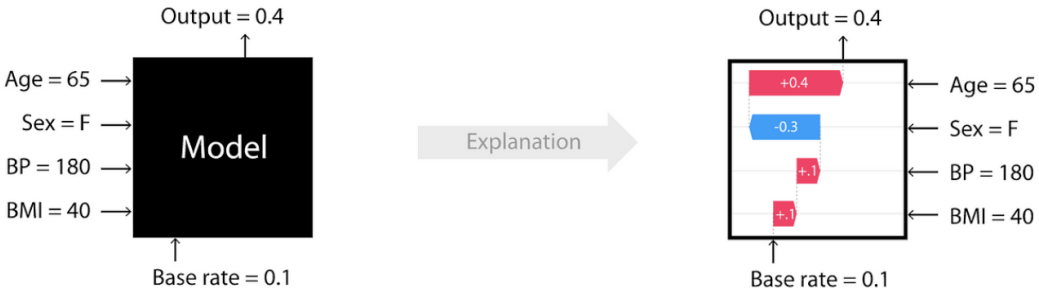


Figure 4.6: SHAP explanations of a black box model

SHAP respects a set of four fairness properties, as detailed in Table 4.1, to ensure the equitable distribution of gains among participants in a collaborative game. This method, presented in [199], serves as an explainability tool that quantifies the impact of individual features on model predictions. In Figure 4.6, we provide an illustrative example of SHAP-based explanation. The figure showcases the contributions of specific feature values, depicted as red and blue polygons, to the final output prediction. The explanation commences from the base rate, indicating the fundamental gain when no features are considered, and progressively highlights the contribution of each feature, culminating in the ultimate output value. SHAP offers the significant advantage of adaptability to a wide range of model types. To further enhance explainability performance, several SHAP variants have been introduced, each optimized for specific types of black-box models. For example, *LinearSHAP* is tailored for linear models, *TreeSHAP* is optimized for tree-based models, *KernelShap* utilizes specially-weighted local linear regression to estimate SHAP values for any model, *DeepSHAP* provides a high-speed approximation for SHAP values in deep learning models, and *Expected Gradients* is designed for neural networks, drawing inspiration from Integrated Gradients. Another notable advantage of SHAP is its ability to generate both lo-

cal and global explanations from individual predictions. However, it is important to mention that with a large number of features, Shapley values become more complex to compute.

4.5 XAI for speech system decision

Most studies in ASpR literature concentrate predominantly on enhancing performance, with only few works focusing on interpretability and explainability of the information embedded within the speaker embeddings. For instance, the work in [201] applied a Gradient-based visualisation method, namely Guided Backpropagation [194], on a CNN-based model, taking as input a raw waveform for speaker identification task. This technique estimates the relevance of the signal to quantify the contribution of each input sample. Recently, the work in [202] proposed the use of three visualisation methods based on class activation map (CAM) [195] to localize the important regions in spectrogram for speakers identification. It shows a saliency map that demonstrates the important regions in the spectrogram used by a ResNet34 with squeeze-and-excitation model [68] to identify a particular class of speaker.

The exploration of model explainability and interpretability have been more emphasized in the general field of speech recognition. The work in [203] developed a bi-directional Gate Recurrent Units based speech recognition model on which the layer-wise relevance propagation (LRP) [196] is applied to explain the recognition task. The explanation provided by this work is presented by the relevance of some parts of the audio sentence into the prediction of the phoneme recognition task. In the same direction, [204, 205] used Shapley values [199] to retain the most important acoustic frames influencing an automatic speech recognition task.

For other different tasks, the work in [206] recently proposed employing post-hoc model-specific methods, including Taylor decomposition [207], LRP, and Integrated gradients [192, 193], to offer interpretations for the deepfake audio detection task. The study focused on a CNN-LSTM model, using a Mel spectrogram as input. Explainable methods were employed to analyze attribution scores calculated based on input energy, that distinguish between deepfake and real voices. In [208], the implementation of LRP was compared across two tasks: digit classification and gender classification using speech data. Each task utilized AlexNet [209] for training, with input being either raw waveforms or spectrograms. This study indicates that specific patterns in the input play a crucial role in the classification of both tasks, using both types of inputs. For explaining speech enhancement task, the authors in [210] proposed an explanation of the predictions of speech enhancement model using DeepShap method. The idea was to figure out which time-frequency bins of the input spectrogram of a noisy signal are used by the DNN model to predict the mask.

4.6 Challenges

In the previous sections, we highlighted the growing interest in interpretability across multiple domains and showcased the diverse array of methods that mark substantial progress in this field. However, it is important to mention that despite these advancements, challenges persist, complicating the task of interpretability in AI models.

- **Trade-off Interpretability/Accuracy**: This arises from the fact that achieving higher accuracy often requires using complex, highly optimized models that are capable of capturing intricate patterns in complex data. While these models excel at prediction tasks, they tend to be less interpretable due to their intricate inner workings. Balancing accuracy and interpretability is a challenge because it forces a choice between two crucial aspects of machine learning [181]. This trade-off dilemma is very dependent of the application in question whether it is worthy to sacrifice performance for the sake of interpretability or not.

- **No formalism of interpretability**: The lack of a universally accepted definition and clear distinction between interpretability and explainability in the field underscores a fundamental challenge [137]. This ambiguity is not only limited to definitions but also extends to the absence of standardized metrics to evaluate the quality and effectiveness of explanations and interpretations provided by XAI methods[181]. This presents a double-edged issue, as without a consensus on what interpretability and explainability mean, it becomes challenging to develop a comprehensive framework for measuring the value and reliability of the generated explanations [145].

- **Data and algorithmic complexity**: The variability within data, intricate feature interactions, resource constraints, limited annotated data, and inherent biases in real-world datasets pose substantial challenges in generating clear and reliable explanations. Consequently, the complexity of AI algorithms employed to handle this data variability further compounds the difficulty of interpretation, making it almost impractical in certain cases.

4.7 Conclusion

In this chapter, we provided an overview of XAI methods within the existing literature, that would serve as a justification for our choices in the remaining of this thesis. We also underscored the critical need for interpretability/explainability in various high-risk domains, including banking, healthcare, and forensics. Moreover, we have drawn attention to the terminology ambiguity that often surrounds the definitions and distinctions between interpretability and explainability in the literature. Additionally, we provided an overview of the limited existing works on XAI in the speech domain. Finally, we acknowledged the inherent challenges of achieving interpretability/explainability and emphasized that it often comes with associated costs.

Although we have used interpretability and explainability interchangeably thus far, we now intend to establish a clear differentiation between them for the remainder of this thesis. For the definition of interpretability we are more aligned with the third category mentioned in §.4.3.

> Interpretability is **" The human ability to understand the model behavior and describe its decisions in understandable way"**.

As an inspiration of many definitions from literature, we precise that:

> Explainability is **"Using methods to describe the output of a black box model in terms of input features either by going through its internal functioning or by using inherently interpretable models that mimic its behavior"**.

We believe that interpretability and explainability are mutually reinforcing concepts. Even in cases where a model is inherently interpretable, the incorporation of explainability can offer additional insights, as highlighted in [182]. Nevertheless, it is crucial to highlight that explanations come with their own costs. They require time and effort, and certain explainability models may introduce further complexity. Thus, the utility of explanations must be carefully balanced against the resources needed to produce them and the necessity of providing them.

In the context of this thesis, we specifically focus on the forensic field as a distinct high-risk application of automatic speaker recognition. Our objective is to explain the decision-making process of our DNN-based ASpR system within this forensic context.

Part II

Proposed solution and contributions

Chapter 5

Inspiration and proposed methodology

5.1 Introduction 60
5.2 Drawing inspiration from DNA: Caution is required! 60
5.2.1 Biological traits information 61
5.2.2 DNA individualisation 62
5.2.3 Misleading phenomena 64
5.2.4 Speech is not DNA! 65
5.3 Proposed methodology 66
5.3.1 Assumptions 66
5.3.2 An overview of the approach 67
5.4 Positioning in the interpretability/explainability dilemma . . . 68
5.5 Conclusion 70

In the previous part, we presented the fundamentals of ASpR systems, the approaches and the advances. Then, we particularly focused on the forensic application of ASpR. Afterwards, we showed the importance of AI models interpretability in such critical field. In this chapter, we introduce our three-steps approach, that aims to build an interpretable and explainable ASpR system suitable for forensic applications. Before delving into the steps of our approach, we begin by exploring the original DNA concept that serves as the inspiration for our work.

5.1 Introduction

Within the forensic context, individualisation or also *individualisation* by DNA is the well-established framework in forensic examination, contributing to the resolution of thousands of crimes. Apart from the challenges associated with the circumstances of a crime, DNA evidence is known for its straightforward interpretation, both by forensic practitioners and the court, thanks to its transparent and easily comprehensible process. Consequently, this framework served as the first inspiration of inference of identity of source with other types of evidence.

In this chapter, we draw a particular inspiration from a simplified DNA forensic individualisation process to build the basics of our approach. This is done while emphasizing on precautions related to the difference between DNA and speech material. Afterward, we introduce our three-steps approach, clarify its assumptions and terminology. Finally, we position our work within the context of interpretability and explainability dilemma.

5.2 Drawing inspiration from DNA: Caution is required!

Deoxyribonucleic acid (i.e. DNA) is the molecule that carries all genetic information about our organism as well as its functioning. In other words, DNA encodes our identity. While we acknowledge the complex nature of DNA, we choose to simplify the concepts for the sake of clarity. In this section, we particularly define some concepts and terminology related to DNA. Following that, we provide a simplified explanation of the individualisation process by DNA in forensics, highlighting some misleading phenomena related to it. Throughout, we present some reflections and we engage in a discussion that explores a potential analogy to the field of speech. This discussion is followed by some research questions that we aim to address through the three-steps of our proposed solution. Finally, we point out that the analogy between DNA and speech is never straightforward and it needs a lot of precautions.

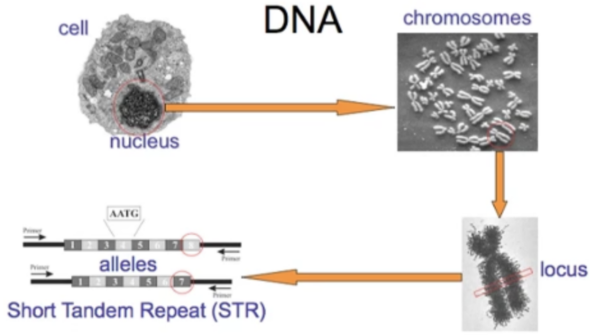


Figure 5.1: Process of alleles information extraction from DNA[1]

[1] https://www.cybgen.com/information/courses/2010/DUCLE/Perlin_DNA_Identification_for_Lawyers_CLE/page.shtml

5.2.1 Biological traits information

As illustrated in Figure.5.1, in each cell's nucleus, our DNA is packaged into 23 pairs of chromosomes including sex chromosomes. In each chromosome, we have loci where the *genes* are encoded. For each gene, we have a pair of *alleles* localised in the same locus originating each from one of our parents. Genes encode physical characteristics (i.e. color of the eyes) and alleles are the different forms of the same trait (i.e. brown).

DNA information is mainly represented by two terms: the genotype and the phenotype as illustrated in Figure.5.2:

- The *genotype* is the set of alleles of an individual for a given gene.
- The *phenotype* is all of its observable characteristics or physical traits which are influenced and determined by both its genotype and the environment factors. The phenotype consists only of visible and expressed traits of a gene.

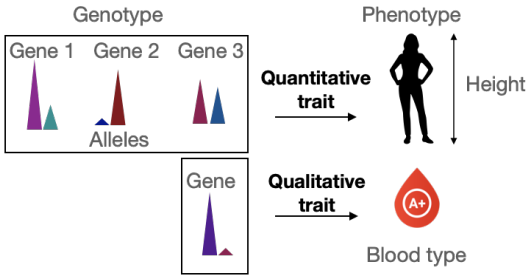


Figure 5.2: Quantitative and qualitative biological traits

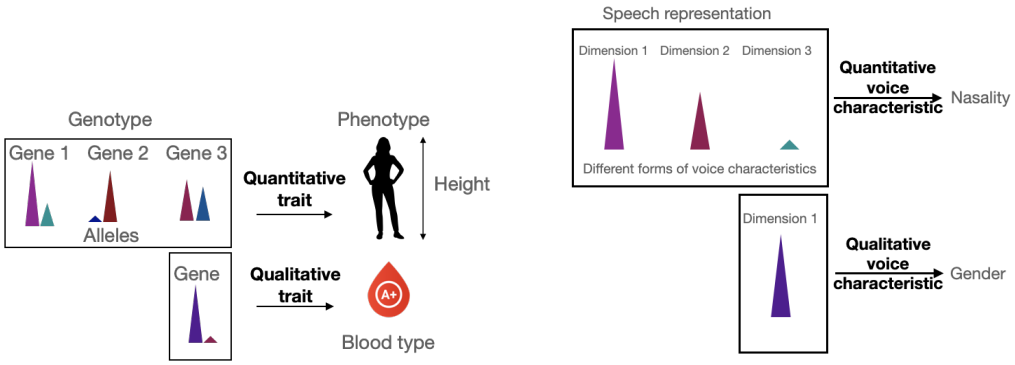
A *biological trait* is an individual characteristic determined by the genotype like hair color, eyes color, height, blood type...etc. One characteristic may have different forms expressed by the variations in the gene that is controlled by the pair of alleles. A trait could be either *qualitative* or *quantitative* as described in Figure.5.2.

- Qualitative traits are controlled by a single gene and they are categorical such as blood type.
- Quantitative traits tend to be more complex and they are usually affected by the environment conditions or controlled by multiple genes in complex interaction between each others like the Height.

Reflection...

These definitions served as inspiration for our work, leading us to draw a specific analogy between speech and biological traits, as illustrated in Figure.5.3.

- What if each gene locus corresponds to a dimension within a specific speaker embedding, encoding a distinct form of a speaker voice characteristic?
- What if alleles correspond to the forms of the voice characteristics encoded inherently by a combination of acoustic, phonetic and phonemic parameters?
- Voice characteristics could be also either quantitative such as nasality that needs many parameters to be encoded or qualitative like sex that is mostly (and could be) determined by the fundamental frequency (i.e. F0).



(a) Quantitative and qualitative biological traits

(b) Quantitative and qualitative voice characteristic

Figure 5.3: DNA analogy to voice characteristic

RQ1: Thus, is it feasible to construct a speech representation, where each dimension represents a form of a specific voice characteristic?

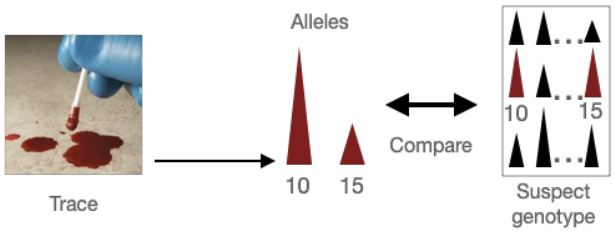
This is a central research question (RQ) that we aim to address in this thesis. However, it is essential to acknowledge that, in contrast to genes, which are predetermined and directly linked to biological traits[2], the dimensions within a speech representation are generally not directly linked to specific forms of voice characteristic neither identified by predefined phonetic and acoustic parameters.

5.2.2 DNA individualisation

Although aware of the challenges associated to DNA individualisation, it is important to note that we are focusing here on an idealized scenario to provide a simplified description. In the process of DNA individualisation, a trace extracted from a crime scene is amplified in the laboratory and transformed to allele data. DNA molecules are amplified during the polymerase chain reaction (PCR) thermocycling process (Figure.5.1). First the alleles are detected by capillary electrophoresis and then the signal is denoised to give the peaks representing the alleles and their height provides a semi-quantiative information on the relative abundance of DNA of the source. A specific

[2]For which a structural physical reality exists

number of peaks is selected as the result of applying a threshold on the amplified alleles to determine the genotype. This is generally applied to reduce the impact of trace contamination and environment factors.



$$LR(\text{Trace, suspect}) = LR_{10} * LR_{15}$$

Figure 5.4: DNA individualisation

The DNA trace genotype is composed of a set of alleles localized at predefined loci [211, 212]. It is compared to the suspect genotype, as illustrated in Figure.5.4. A likelihood ratio is therefore calculated to evaluate the value of evidence given the prosecution H_p (i.e. there is a match between the trace and the suspect), and the defence hypotheses H_d (i.e. the trace matches someone else in the population by coincidence). In DNA, a partial LR is calculated for each locus independently, using the presence or absence of alleles in both parts of the comparison under the two alternative hypotheses [213]. Thanks to the independence between the predefined loci involved in the comparison, the global LR is therefore obtained as the product of these partial LRs (Figure.5.4). This method of quantifying the evidence enables the comprehension of the significance of each locus linked to a specific gene and the factors influencing the evidence's values. It offers simplicity in interpretation for the court and clarity in explanation for the forensic practitioner [214, 215, 119].

Reflection...

The idea of calculating the final LR as the product of the partial LRs, each related to a specific allele and calculated based on the presence or absence of the allele in the genotype, seems very intuitive. This formulation is generally straightforward for the forensic practitioner and easily interpretable by the court.

Indeed, the independence assumption between the predefined loci is the reason that made this LR computation feasible without lose of information. Returning to the earlier analogy, in speech context, it is important to acknowledge that dimensions are generally not independent in the speech representation, unless applying some constraints to push it this way. Constraining independence between dimensions is a very challenging task in speech, because dimensions are often encoding low-level features in a complex interaction between each others. This constraint should be carefully considered and not neglected.

RQ2: In continuation of the previous question, what if we represent a speech sample by a vector indicating the absence or the presence of a certain voice charac-

teristic, ensuring independence between dimensions? Then, we calculate the LR in a DNA-like fashion?

5.2.3 Misleading phenomena

While the suspect genotype involves the whole profile of the suspect signaling all alleles, the genotype of the trace could be presenting some alleles that were masked, hiden or altered due to extraction conditions and environment factors. This introduced uncertainty is due to the nature of forensic conditions that may lead to insufficient, masked or degraded material. This may lead to two main phenomena [216, 217, 218, 219]: *Drop-in* and *Drop-out*.

Drop-in

It describes the presence of "foreign" allele in the DNA genotype [220]. It is defined as the false detection of an allele. The drop-in phenomenon is typically associated with poor DNA conditions. The drop-in phenomenon occurs when a fragmented DNA molecule contaminates a tube or other consumable that contains a sample extract. This results in the appearance of an extra allele that cannot be attributed to the known suspect genotype [220, 216]. Since drop-in is related to noise and contamination of data. it is shown to be a bit difficult to estimate mathematically [216].

Drop-out

It follows the principle: *"UNSEEN does not mean NOT EXISTED"*. Drop-out is the phenomenon of missing alleles at one or more loci. The reasons behind this disappearance may be a dominant allele randomly fails to PCR amplify, or an existing allele misgenotyped because of factors such as PCR or electrophoresis artifacts or human errors in reading and recording data [221, 218, 217, 222]. Drop-out is estimated by a logistic analysis [223] or by an empirical approach [224].

This uncertainty is involved into the partial LRs estimation of each locus for a better quantification of the evidence [222, 217, 216].

Reflection...

Indeed, the integration of these uncertainty phenomena into the LR calculation enhances the interpretability of the value of evidence and boosts confidence in the quantified information [214]. It is our belief that incorporating both of these phenomena into the context of speech data can be fully justified. In speech, the phenomenon of drop-out could occur in two scenarios: 1) either the voice characteristic is not present due to speech variability, or 2) the voice characteristic is not detected due to insufficiency of data, where the characteristic is related to particular linguistic or phonemic aspect[3]. The phenomenon of drop-in could as well occur in speech due to noise caused

[3]The characteristic is linked to a specific phoneme or a specific group of phonemes which are not part of the recording

by multiple factors such as environment, recording conditions, quality of device, etc. Drawing attention to these two phenomena could be an innovative approach to model the uncertainty in speech data.

RQ3: Building upon our analogy, what about quantifying and incorporating both uncertainty phenomena in the LR calculation of speech evidence?

5.2.4 Speech is not DNA!

Thus far, we draw a distinctive analogy between DNA information and speech characteristics. However, we are aware that this analogy is far from the real world scenario and it is essential to approach it with careful precautions:

1. It is important to mention that DNA scenarios and challenges (e.g. DNA mixture) are significantly more complex than the simplified version described here. Our main intention is to simplify the process for the sake of clarity and for an easier comprehension of the core concept.

2. While acknowledging the intricate interplay that can exist among genes encoding biological traits, it is important to mention that almost all of these genes are predefined and localized in specific loci. This is in contrast to the extraction of speech representation, that does not capture all the aspects of speech, where a single vocal characteristic is encoded by information from numerous abstract dimensions that lack inherent interpretability.

3. Indeed, unlike DNA genotype, which reflects a full profile of an individual, speech representations model specific speech segments with distinct content, background noise, and speaker variability. While a DNA genome comprises all the ground truth information about an individual, the notion of a *speaker profile*, containing all conceivable variabilities, from aging to emotions, health conditions, and background noise remains a myth.

4. It is critical to mention that for DNA, drop-in and drop-out phenomena are occurring in the trace side because it is prone to error, but not in the suspect genotype which presents the ground truth genetic information. In contrary, for speech data these phenomena are probable to happen in both sides, in the trace sample as well as the suspect sample.

Despite these precautions, we firmly believe that the inspiration from biological DNA is innovative and valuable. It can be viewed as the ideal scenario that quantifies and presents the value of evidence in an understandable manner, while retaining only the useful information.

5.3 Proposed methodology

As illustrated in the literature review part of this thesis, the use of automatic DNN-based speaker recognition systems in high-risk applications such as forensic science requires specific consideration. The inherent opacity of these AI-driven systems, employed to assess the value of speech evidence, may raise concerns about potential discrimination bias and provokes ethical questions regarding the reliability of the evidential value. To address the lack of interpretability in DNN-based ASpR systems, both in the general context and more specifically within the field of forensics, we introduce in this section our proposed three-steps solution. We start by presenting the assumptions, building upon the above DNA inspiration, then we give an overview of our proposed approach.

5.3.1 Assumptions

Following the inspiration drawn from DNA individualisation, we establish some assumptions as the foundation for formulating our solution. Figure.5.5 illustrates some of these assumptions in a toy example of an ideal speech representation composed only of three voice characteristics. So in this work we assume that:

- *Assumption 1*: A speech sample is represented by a discrete representation, where each dimension is assumed to encode a specific form of voice characteristic or an *Attribute*. E.g in Figure.5.5 we show three forms of three voice characteristics.

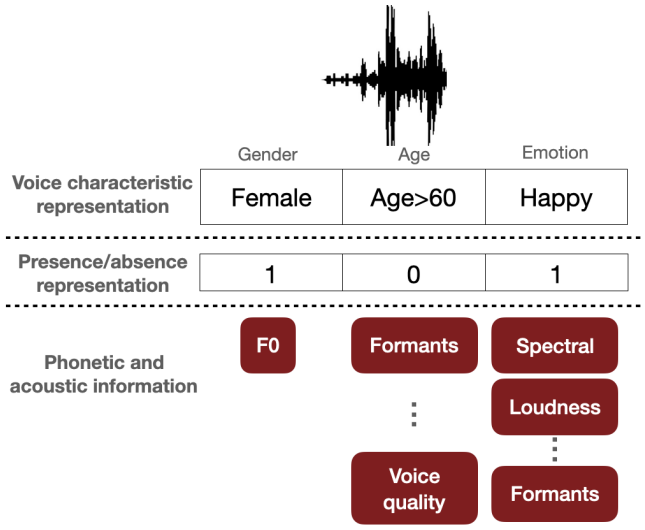


Figure 5.5: An illustration of assumptions of the ideal representation

- *Assumption 2*: A voice attribute is assumed to be shared between a group of speakers.
- *Assumption 3*: Although the number of voice characteristics is undefined, we assume a closed set of a reduced number of voice characteristics.

- *Assumption 4*: Whatever the voice attribute is qualitative (e.g. female in Figure.5.5) or quantitative (e.g. Age=61), we assume them all to be qualitative by converting them to categorical (e.g. category of ages>60 in Figure.5.5).
- *Assumption 5*: In the speech representation, each dimensions is assumed to encode the presence (i.e 1) or absence (i.e 0) of the form of voice attribute (Figure.5.5).
- *Assumption 6*: Each form of a voice attribute encoded in a dimension is assumed as inherently encoded by a combination of acoustic, phonetic and phonemic parameters (Figure.5.5).
- *Assumption 7*: The dimensions of the speech representation are assumed to be highly decorrelated or independent.
- *Assumption 8*: Drop-in and drop-out phenomena are assumed to occur in speech data and are used to quantify uncertainty.

5.3.2 An overview of the approach

In this section, we present our solution for approaching the lack of interpretability and explainability in DNN-based ASpR systems. The primary objective of this solution is to offer an interpretable and explainable system for speaker recognition in general and, more specifically, to enhance the evaluation of the evidence in forensic context. This approach is mainly inspired from the interpretable DNA individualisation process, and built upon the formulated assumptions, while respecting all necessary precautions.

In the following, we provide an overview of our proposed solution, as illustrated in Figure.5.6, which mainly consists of three key steps.

Step 1: Binary and attribute-based speaker embeddings

In this first step, our goal is to model a binary and attribute-based speech representation, following the previously set assumptions. This specific representation is mainly inspired from the concept of DNA genotyping where the absence or presence of an allele is considered for forensic DNA individualisation (as described in §5.2.2). In this step, we aim to represent a speech segment by a binary vector, where each dimension of the vector indicates the presence (i.e. 1) or absence (i.e. 0) of a specific attribute. This representation is based on SOTA ASpR systems and automatically generates attributes. Further description and details about the extraction process as well as the used model are discussed in chapter 6, along with some improvements proposed in chapter 10.

Step 2: Binary and attribute-based likelihood ratio estimation

In this second step, we aim to elaborate an interpretable and explainable scoring process that uses the binary and attribute-based speaker embeddings, to assess the value of

evidence for a forensic speaker recognition task. During a test scoring of a given pair of speech samples, a LR is computed for each attribute independently of the others, referred to as *attribute-LR*. This LR per attribute is computed under the prosecution and defence hypotheses. As illustrated in Figure.5.6, it is based on: 1) the value of the attribute in both sides of the comparison, 2) An estimation of the discrimination of the attribute. and 3) An estimation of uncertainty of the attribute (i.e. drop-in and drop-out). A global LR is then computed as the product of these attribute-LRs, assuming that these attributes are independent. This process is untitled *BA-LR framework*, referring to as *Binary-Attribute-based LR*. This approach is far more informative than an LR derived solely from a similarity score. The resulting LR not only serves as a transparent framework for forensic practitioners to enhance their understanding of the evidence but also acts as an interpretable process for decision-makers, including juries and judges, offering greater insight into the information that describes the value of evidence. Consequently, this would afford more control and flexibility in the process of evidence evaluation. The formulation of this framework as well as the proposed scoring process, inspired from DNA individualisation, are addressed in more details in chapter 7.

Step 3: Attribute explainability

Up to this stage, attributes encoded in the speaker embeddings are issued from a bottom-up process and are not yet identified or described in an understandable manner. In this third step, we propose a methodology that aims to describe these attributes. It is based on a theoretical modeling of three worlds: a real world that represents the speech data, a representation world which is the binary speaker embeddings and an informative world which contains all information concerning speech samples such as acoustic, phonetic and phonemic parameters. The goal is to find a mapping function between the representation world and the informative world to provide an automatic description for attributes. Thus, two attributes descriptions are provided using two types of mapping functions, as described in Figure.5.6: 1) An utterance-level mapping that provides phonetic description of the attribute. 2) A frame-level mapping that describes attributes in terms of localized temporal information, phonemes and phonetic classes. Both approaches aim to enhance the explainability level of the attribute and to prepare it to be more interpretable by an expert. This is further detailed in chapter 8.

5.4 Positioning in the interpretability/explainability dilemma

As discussed in §.4.3, the lack of a unified definition of interpretability and explainability poses a significant challenge. In order to avoid ambiguity, we proposed a definition for each term that we believe are suitable for this work in §.4.7. In the following, we remind these proposed definitions:

- Interpretability: The human ability to understand the model behavior and de-

scribe its decisions in understandable way.

- Explainability: Using methods to describe the output of a black box model in terms of input features either by going through its internal functioning or by using inherently interpretable models that mimic its behavior.

With these definitions in mind, we believe that our three-step solution tackles both, interpretability and explainability. Our goal is to clarify our position, at each step, within the definitions and taxonomy (§4.4) of explainability and interpretability, as described in Figure 5.6.

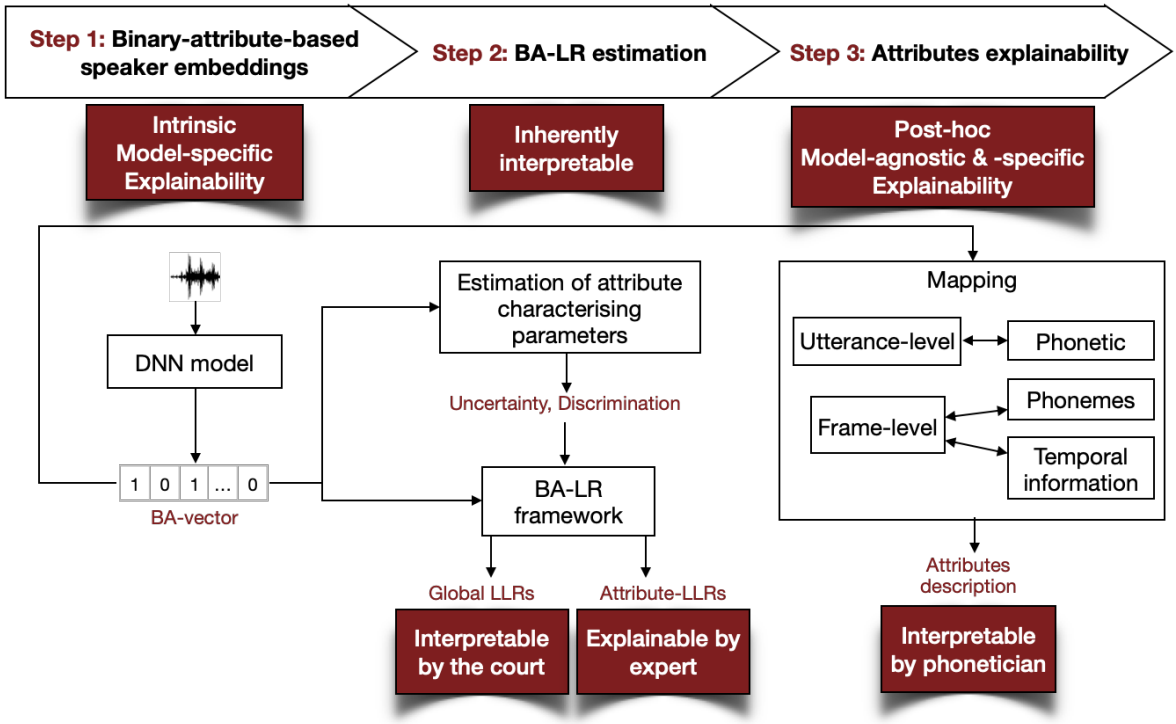


Figure 5.6: Positioning of our approach in the interpretability/explainability dilemma

- In Step 1 of our approach, and according to the established definitions, orienting the speaker model to form binary attributes within speaker embeddings offers explainability. It does not directly yield interpretable representations in an intuitively "understandable way" but rather, it serves as a significant step that generates representations that are ready to explore and more easy to explain. Thus, as described in the proposed taxonomy §4.4, this constraint falls under the categories **Global, Intrinsic and Model-specific explainability**.

- The main goal of Step 2 is to provide a computation of the LR that is both explainable by the forensic expert in terms of contributing factors, and interpretable by the court in terms of evidence quantification. The BA-LR framework is intentionally designed to be **inherently interpretable**, explicitly revealing the

role of each attribute in the final LR calculation. Together with a characterization of the attribute, this enhances the interpretability of this framework for the court, enabling them to more easily comprehend the presented evidence value. Specifically, the attribute-LR values allow the forensic practitioner to explain the contribution of each sub-process to the final decision both locally and globally. At this stage, still both the interpretability and explainability of the value of evidence by both the court and the forensic practitioner is not fully satisfied. Indeed, we have an estimated characterization of each attribute, but yet they are not described or identified in a humanly understandable way. This is addressed in the next step.

- Step 3 aims to describe the attributes deriving from DNN model, where no information about their nature is provided. For this end, we propose a methodology that describes the attribute at two levels: 1) At the utterance-level using a **Global, Post-hoc, Model-Agnostic explainability** method to automatically describe attributes in terms of phonetics. 2) At the frame-level by reversing the DNN's flow to explore the relationship between each attribute and its associated phonemes and temporal information. This approach is considered as **Post-hoc, Model-specific** because it is specific to the DNN. The attribute description provided by this step is more accessible for a forensic phonetician, who can listen to the audio regions, interpret this information and translate it into higher-level voice characteristics more comprehensible for the court.

5.5 Conclusion

Even thought the inspiration from DNA might be dangerous, we firmly believe that it sparked a novel and innovative idea. While carefully approaching this inspiration, we aim to create a new direction and elaborate the core concept of our solution. In the rest of this thesis, we further detail the proposed three-step, presenting each step in a dedicated chapter.

Chapter 6

Step 1: Binary and Attribute-based modelling of speaker embeddings

6.1 Introduction 72
6.2 Related work: Binary speaker embedding 72
6.3 Binary-Attribute-based modelling 73
6.3.1 Requirements 73
6.3.2 Proposed model: ResNet with thresholding 74
6.4 Experiments and results 76
6.4.1 Setup 77
6.4.2 BA-vectors analysis 78
6.4.3 Measuring attributes correlation and dependence . . . 79
6.4.4 Speaker recognition performance 81
6.5 Discussion 82

This chapter describes the architecture of the DNN-based extractor employed in Step 1 of our solution. This DNN model is designed to extract binary-attribute-based speaker embeddings. In this chapter, we present an initial version of this extractor which serves as the foundation for the subsequent steps of our work.

6.1 Introduction

As seen previously, the analogy between speaker recognition and DNA identification is far from straightforward. Nevertheless, this inspiration can serve as a motivation to create a more interpretable modelling of speaker embeddings, driven by the presence or absence of a set of attributes. In contrast to the dense and unstructured nature of continuous DNN-based speaker representations, such binary representation may offer a more efficient means of organizing and condensing diverse speaker information into distinct dimensions, thereby simplifying the handling of encoded information.

In this chapter, we conduct an overview of the existing methods in the literature that produce binary speaker embeddings. Then, we precise requirements for this representation as well as the proposed DNN-based extractor. After, we conduct some experimental analysis to evaluate the resulting representation with regard to the requirements and ASpR performance. Finally, we discuss the results and summarize our conclusions and perspectives.

6.2 Related work: Binary speaker embedding

In the literature, a limited number of works only focus on representing a speech sample with a binary representation [225, 226]. Most of the works are specifically dedicated to generate binary speaker embeddings for speaker recognition task. We categorize these works into three main groups based on their objective: 1) to model speaker specific discriminant information, 2) to preserve privacy and enhance the security of speaker information in the embedding, and 3) to reduce both time and computation costs.

Regarding the first group, the work in [227, 228] proposed an approach that moves from a continuous probabilistic space based on GMM-UBM to a discrete, binary space, able to handle directly the speaker discriminant information. This process leverages an UBM to partition the acoustic space into distinct regions. Within each region, a set of Gaussian models is employed to highlight speaker-specific characteristics. Consequently, each acoustic frame is transformed into the discriminative binary space, where it activates or deactivates various specificities, creating a binary vector. The work in [229] builds upon the same approach while exploring both frame- and segment-level speaker specific information.

The second group uses binary speaker embedding for security purposes, referred to as *Secure binary embedding* (SBE) [230]. The work in [231, 232] builds upon the i-vector system by computing SBE hashes of i-vectors. It converts real-valued i-vectors to bit sequences which represent the SBE hashes in such a way that when the Euclidean distance between two i-vectors falls below a predefined threshold, the Hamming distance between their respective hashes is directly related to the Euclidean distance between the i-vectors. But, if the Euclidean distance surpasses this threshold, the hashes fail to provide meaningful information about the actual distance between the two i-vectors.

Other researchers took benefits from binary vector to reduce computational costs.

They generally use *locality-sensitive hashing* (LSH)-based technique for performing fast similarity searches. This technique consists in computing randomized hash functions that encourage a high probability of collision for similar vectors. For instance, [233] presented a binarization approach within the GMM-UBM framework that efficiently search large populations of speakers using kernel LSH [234]. The work in [235] converts i-vectors to binary vectors or codes by a hash function. The binary codes are obtained by both a LSH using a set of random hash functions and a Hamming distance learning approach, where the hash function is trained using variable-sized blocks, independently projecting each dimension of the original i-vectors into variable-sized binary codes. More recently, the work in [236] proposed an ordered binary auto-encoder for speaker recognition that sorts the dimensions of the embedding vector, the x-vector, and converts the sorted vectors to binary codes using Bernoulli sampling. This is shown to reduce memory storage and speed up retrieval tasks.

Despite the scarcity of research on this topic, representing a speaker embedding with a binary vector offers numerous advantages. These include handling high dimensionality while maintaining a compact representation, concealing speaker information for security purposes, and guaranteeing reduced memory storage, minimized computational and time costs, and expedited retrieval tasks.

Clearly, our goal is not to generate a binary code through hashing functions where dimensions lack meaning beyond quantization. Instead, we aim to create a binary-attribute-based speaker embedding that encodes speaker-specific information within a speech recording. This aligns closely with the first group of methods [227, 228, 229]. Although these methods are based on statistical GMM-UBM models, the idea behind their binarization approach is intuitive and inspiring for our requirements.

6.3 Binary-Attribute-based modelling

In this step, we introduce a binary-attribute-based modelling that encourages speaker embeddings to concentrate speaker-specific information into distinct dimensions, referred to as attributes. This is tackled following a set of requirements and aligning with DNN-based ASpR architectures. In this section, we clarify these requirements. Following that, we introduce our proposal and provide rationales behind our selection of the architectural design.

6.3.1 Requirements

We start by defining some important requirements for the proposed modelling of speaker embeddings, as follows:

- **Fixed-sized binary vector**: a fixed-length vector of binary attributes. Each dimension indicates whether the corresponding voice attribute is present (i.e. 1) or absent (i.e. 0).

- **Shared attributes**: an attribute is shared by a specific group of speakers. Attributes should follow the behavior of voice characteristics where some are shared by few speakers (rare), others are shared by half of the speakers, and others shared by most of the speakers (typical).

- **Independence**: attributes are assumed to be independent or highly decorrelated.

6.3.2 Proposed model: ResNet with thresholding

With these requirements in mind, we propose for this step an initial solution built upon the ResNet DNN-based extractor, outlined in §2.3.3. In this section, we present the motivation behind this extractor and the constraints applied to push the desired representation. Then, we set some notations that we find useful for this work.

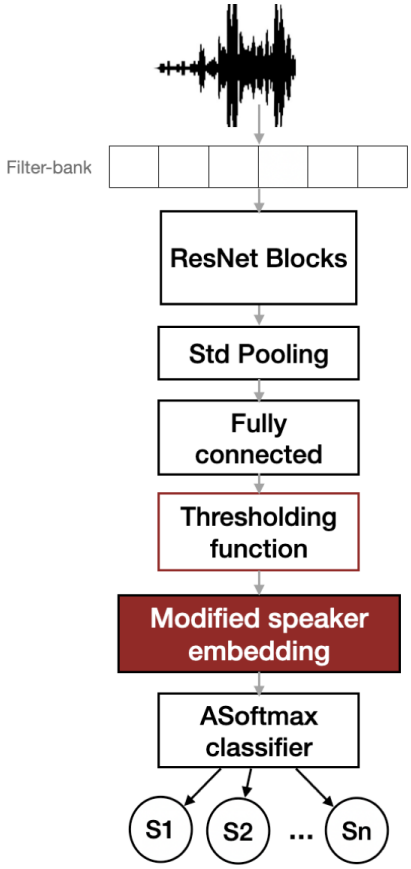


Figure 6.1: ResNet training with thresholding function

The idea is to incorporate a thresholding function into the DNN model to promote a substantial contrast in the activation levels among different dimensions (i.e. neurons) of the speaker embedding. The objective is to guide the network to prioritize certain dimensions during training, promoting high activation in these dimensions, while allowing the others to remain weakly activated or ignored. This thresholding function

is an activation function that is added at the utterance-level of the network after the fully connected layer as illustrated in Figure 6.1.

This modified model is therefore trained for a speaker classification task. During training, the model pushes some neurons to be more activated than others for each speaker class, referred to as *speaker attributes*. This way and given the fixed-length of the embedding, each attribute would be certainly shared by a subset of speakers. After the training process, the extracted speaker embeddings remain non-binary. To address this, we apply a post-training threshold, which converts very low values (i.e., <1e-4) to 0 and sets positive values to 1. This way we obtain binary-attribute-based speaker embeddings.

Why ResNet ?

The initial choice of ResNet is mainly driven by the observation that it offers a satisfied compromise between complexity and performance (Table 2.2), in our point of view, compared to TDNN that is less accurate and ECAPA-TDNN that is much more complex to follow for the next steps of the approach. However, it is important to note that involving such thresholding function into the DNN model is totally independent of the architecture and could be applied to any other ASpR network.

Which thresholding function?

ReLU function is the most used activation in DNN models training. It is linear for positive values and zero in the origin and over the negative values as illustrated in Figure 6.2 in red. Thus, when the input becomes zero or negative, the gradient of the function becomes zero and as a consequence it will not perform the backpropagation operation. This is called "The dying ReLU". To prevent such phenomenon during training, a smooth variant of ReLU, referred to as *Softplus function* [237], is shown to be differentiable around zero allowing more small negative values to be activated as shown in Figure 6.2. The vectors deriving from both functions (refer to Figure 6.1) after training are referred to as *activation vectors*. To achieve binary speaker embeddings, a *post-training thresholding* function is employed on activation vectors, as illustrated in Figure 6.2, where green threshold is dedicated to Softplus function, red threshold referred to ReLU, x is the input to the activation function and $f(x)$ is the activation values. The post-training thresholding function is expressed in the following equation as:

$$\begin{cases} 0 & f(x) < 1e^{-4} \\ 1 & Else \end{cases} \tag{6.1}$$

Clearly, the post-training threshold is applied to obtain exactly zero values. This is because during training, values do not become exactly zeros, neither for ReLU nor Softplus. Instead, we obtain vectors with very small values close to 0, indicating almost

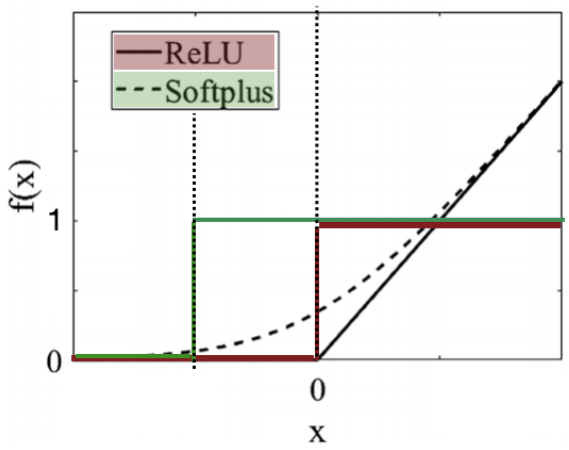


Figure 6.2: ReLU and Softplus activation functions, along with post-hoc thresholding functions.

null activation. This phenomenon is frequent when no regularization is applied to the training of the network.

Notations

For clarity reasons, we unify the notations used in our approach by denoting the following:

- *BA*: A given Binary Attribute.
- *BA-extractor*: The ResNet model modified with a thresholding function during training.
- *BA-vector*: The vector of binary attributes, referred to as binary-attribute-based speaker embedding.
- *ReLU-vector*: The vector of activations derived from ReLU.
- *Softplus-vector*: The vector of activations derived from Softplus.
- *Neurons*: Refers to the dimensions of ReLU- or Softplus-vectors.
- *Attributes*: Refers to the dimensions of the BA-vector.

6.4 Experiments and results

In this experimental section, we begin by introducing the experimental framework. Following this, we conduct a comparative analysis between ReLU and Softplus activations. This analysis involves examining the distribution of 0s in attributes for both

ReLU and Softplus vectors and measuring the correlation and dependence between dimensions (i.e. neurons) in both types of vectors. Finally, we evaluate the performance of the extracted BA-vectors in a speaker verification task.

6.4.1 Setup

We firstly outline the experimental setup for this step, which involves a description of the data sets used and the evaluation protocol designed for speaker verification task. Subsequently, we specify some details regarding the extractor configuration.

Data sets and protocol

This step is mainly performed using VoxCeleb data set [57, 39], a commonly used corpus frequently used in ASpR systems available in two versions: VoxCeleb 1 & 2. It consists of short clips of human speech, extracted from interviews with celebrities in YouTube. The dataset is diverse, with speakers from 145 different nationalities, offering a wide range of accents, ages, ethnicities, and languages. A description[1] of both data sets is provided as follows:

- VoxCeleb2 [63]: It is a vast speaker recognition dataset consisting of more than a million utterances from over 6,000 speakers. It mimics a real-world dataset with various types of noise like laughter, cross-talk, music, and other environmental sounds. Table 6.1 summarizes the number of speakers and speech extracts in the development set which we use in this work.
- VoxCeleb1 [238]: consists of more than 150,000 utterances from 1251 celebrities. It is almost balanced in term of gender, with 55% male and 45% female.

Table 6.1: Data set and protocol description

	VoxCeleb2	***VoxCeleb1***
# of speakers	5,994	1,251
# of extracts	1,021,175*5[1]	153,516
# Test pairs		56,295*2[2]

[1] 1 original and 4 augmented versions per extract
[2] 1 for target and 1 for non-target trials

VoxCeleb1 and VoxCeleb2 datasets do not share any speakers in common. In this work, we use VoxCeleb2 dataset for training the DNN extractor. Additionally, we apply data augmentation [46] using MUSAN dataset[2], such as adding babble, reverberation, music, and incorporating noise to introduce more variability and improve the

[1] https://www.robots.ox.ac.uk/ vgg/data/voxceleb/
[2] http://www.openslr.org

DNN model's robustness (refer to Table 6.1). VoxCeleb1 primarily serves as a test set to evaluate our DNN embeddings based on the specified requirements and speaker verification performance.

Using this test set, we construct the comparison pairs useful for speaker verification task following the protocol outlined below; we create a set of comparison pairs using the test set, selecting the first ten utterances for each of the 1251 speakers in VoxCeleb1 to form the target (i.e. same-speaker) and non-target (i.e. different-speakers) pairs. This yields a total of 56,295 target comparison pairs, with 45 pairs for each speaker as shown in Table 6.1. To maintain a balanced dataset, an equal number of 56,295 non-target pairs are randomly selected.

BA-extractor configuration

The BA-extractor is mainly based on the baseline ResNet architecture [59, 57] (refer to §2.3.3 for more information). It takes as input 61-filterbank extracted with Kaldi recipe [239]. The extraction of filterbank values follows the same process described in §.2.3.1 and §.2.3.1. The ResNet comprises of four consecutive residual CNN blocks, followed by a standard deviation pooling layer (Std) and a fully connected layer with batch normalization, which ultimately yields 256-dimensional speaker embeddings. Our thresholding process is applied to this fully connected layer, resulting in the modified speaker embeddings.

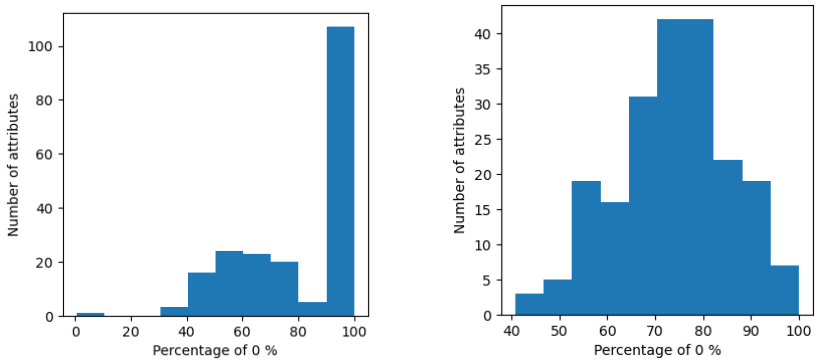
For this experiment, we deploy three models: a baseline ResNet model and two different modified models, each featuring a distinct activation function. In the first modified model, we introduce a ReLU activation solely to the fully connected layer, preserving the baseline architecture intact. In the second model, we make specific alterations by removing the batch normalization associated with the fully connected layer to prevent the spread of speaker information across neurons during training. Subsequently, we incorporate the Softplus activation function. All models are optimized for a speaker classification task using the ASoftmax objective [83].

6.4.2 BA-vectors analysis

In this section, we carry out analyses of the BA-vectors issued from activation functions extracted from the test set. This analysis would allow us to study closely the behavior of ReLU and Softplus activations with respect to the resulting embeddings.

Figure 6.3 shows the distribution of zeros in BA-vectors derived from both ReLU- and Softplus-vectors of the test set. It is important to mention that a phenomenon known as 'dead neurons' is observed in the BA-vectors, where specific neurons remain inactive across all speech excerpts (i.e. 100% of zeros). This phenomenon substantially reduces the total number of attributes. In our case, this led to a reduction in the number of attributes derived from both activation functions, with 256 attributes reduced to 197 and 205 for ReLU- and Softplus-vectors, respectively. Furthermore, a important portion of the attributes (i.e., over 100) from ReLU-vectors exhibit a high percentage

of zeros across speech excerpts (refer to Figure 6.3a). Both phenomena may lead to a substantial loss of information within the corresponding BA-vectors. However, this is not the case for Softplus-vectors, where the distribution of zeros within attributes demonstrates a more balanced pattern compared to ReLU-vectors (see Figure 6.3b). This difference may be attributed to the Softplus function's ability to smoothly accommodate certain small negative values while maintaining their activation, in contrast to the ReLU function, which deactivates all negative values.



(a) Binary attributes from ReLU-vectors

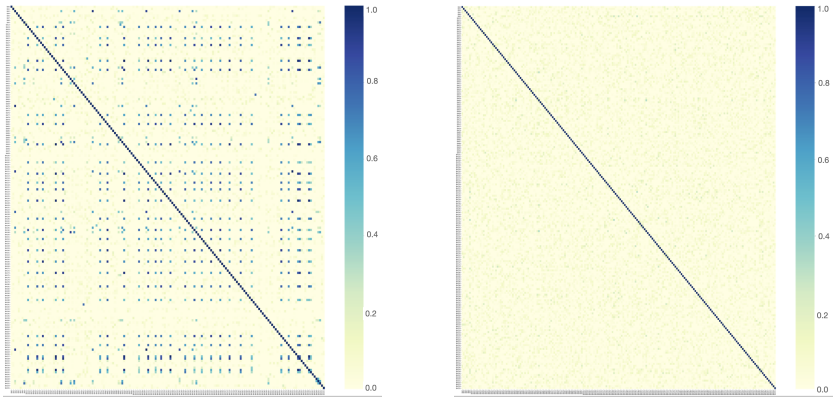
(b) Binary attributes from Softplus-vectors

Figure 6.3: Distribution of 0s in the BA-vectors

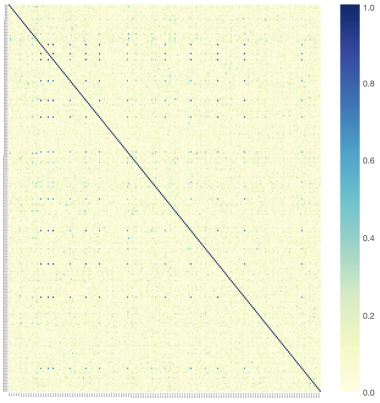
6.4.3 Measuring attributes correlation and dependence

Indeed, in this BA-extractor, no constraint is applied to encourage the independence assumption. Thus, it is not theoretically guaranteed. In this section, we aim to experimentally evaluate its level of compliance. For comparison reasons, we show in Figure 6.4 three Pearson correlation matrices, that measure the linear correlation between the dimensions of the x-vector and between the neurons of both ReLU and Softplus-vectors. To have more closer look into correlation values, we show their respective distributions through boxplots in Figure 6.5. Upon comparing the three matrices, it is clear that the baseline x-vectors exhibit already relatively low correlation between dimensions. Although there are a few high correlations among specific pairs (refer to Figure 6.5c), the overall correlation matrix predominantly reveals moderate values, as depicted in Figure 6.4c. Unlikely, ReLU-vectors exhibit a more pronounced contrast in correlations, with notably higher correlations observed between certain pairs of dimensions, while showing notably lower correlations among others as shown in Figure 6.4a. On the other hand, the Softplus-vectors exhibit notably weak correlations, with almost values falling within the range of -0.2 to 0.3, as illustrated in Figure 6.5c.

To further illustrate this decorrelation, we employs another measure, the *Mutual Information* (MI) [240], that not only considers linear correlations but also takes into account independence which is a stronger statement than decorrelation. It is estimated

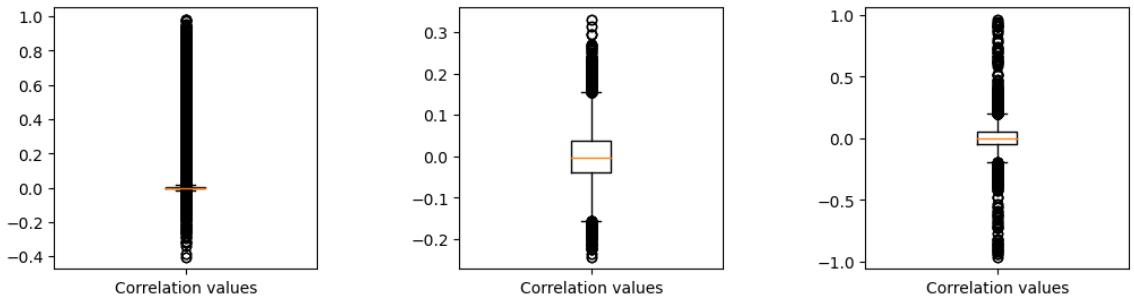


(a) ReLU activations (b) Softplus activations



(c) X-vectors dimensions

Figure 6.4: Pearson correlation between neurons activation



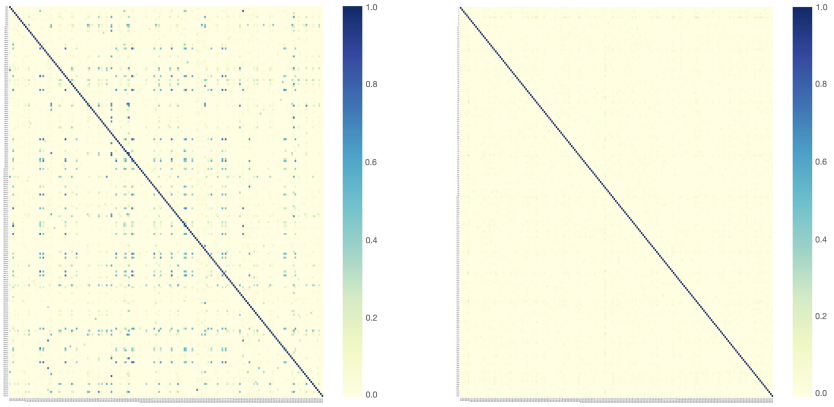
(a) ReLU correlations (b) Softplus correlations (c) X-vectors correlations

Figure 6.5: Distribution of correlation values in boxplots

as follows:

$$MI(x,y) = \sum_{x}\sum_{y} P(x,y) \cdot \log(\frac{P(x,y)}{P(x) \cdot P(y)}) \tag{6.2}$$

Where $P(x,y)$ is the joint probability and $P(x)$ and $P(y)$ are the marginal probabilities. Figure 6.6 shows the mutual information calculated between binary attributes deriving from both ReLU- and Softplus-vectors. As can be noticed, the mutual information between binary attributes of Softplus-vectors reveals a notably low level of dependence between attributes (see Figure 6.6b), in contrast to ReLU, where certain attribute pairs exhibit a considerably higher degree of mutual dependence (refer to Figure 6.6a).



(a) ReLU binary attributes (b) Softplus binary attributes

Figure 6.6: Mutual information between binary attributes

6.4.4 Speaker recognition performance

Table 6.2 presents the results of a speaker verification process employing the baseline x-vector model, ReLU model, and Softplus model. The results report the performance in terms of EER (§2.4.1) and $C_{llrmin/act}$[3] (§3.3.3). The evaluation of the three systems is conducted using cosine similarity scores (§2.3.4), which are calculated based on comparison pairs composed from VoxCeleb1 following the experimental protocol set earlier.

Table 6.2: Performance comparison in terms of EER and C_{llr} of the three systems on VoxCeleb1 using cosine similarity.

	X-vectors	**ReLU-vectors**	**BA-vectors**	**Softplus-vectors**	**BA-vectors**
# of dim	256(float32)	197(float32)	197(bit)	205(float32)	205(bit)
EER	1.37%	3.84%	5.82%	3.38%	3.42%
Cllr$_{min/act}$	0.057/0.81	0.14/0.89	0.21/0.95	0.13/0.86	0.13/0.86

As can be noticed in Table 6.2, the overall performance of both ReLU and Softplus-vectors has shown a decrease compared to the baseline x-vector. This performance

[3]The minimum C_{ll}/ the actual C_{ll}

decline is manifested by approximately 2% average absolute increase in EER for both activation vectors. This loss can be explained by the attenuation and suppression of some dimensions, leading to a more compact binary-attribute-based embeddings. This reduction in performance is more pronounced in the case of ReLU-vectors than in Softplus-vectors. One potential explanation of this gap may be attributed to the smooth behavior of Softplus, which allows for the acceptance of certain small negative values as 1. When comparing the BA-vectors, one can notice that the conversion of ReLU-vectors to BA-vectors results in a supplementary increase in EER, with nearly 2%. In contrast, binarizing Softplus-vectors, interestingly, maintains almost the same performance, with negligible reduction of 0.04%.

6.5 Discussion

In this chapter, we proposed a first version of the binary-attribute-based extractor. We showed that extracting speaker embeddings that respect our requirements is a challenging task. The idea to incorporate an activation function at the speaker embedding level seems to be an appropriate trick, which reshapes the embeddings, orienting them towards binarization. Furthermore, we believe that integrating this activation function during the model's training process effectively concentrates speaker-specific information into selected neurons. This would guide the most of the flow of information within the DNN model layers towards these targeted neurons. Given that the training is optimized for the speaker classification task, these neurons would subsequently function as shared attributes across groups of speakers.

Although not directly involving the independence constraint into the DNN model training, we demonstrated that this constraint is satisfied experimentally through correlation and dependence measures. Our findings illustrate that the inclusion of activation function at the speaker embedding level is advantageous with respect to our requirements. It is thought to efficiently alleviate information correlation between attributes, resulting in pronounced decorrelation and increased levels of independence. In particular, the utilization of the Softplus activation demonstrates an enhanced ability to decorrelate the activation vector neurons, thereby promoting attribute independence. Furthermore, it has shown greater stability in activations, with the added benefit that converting these activations to binary attributes does not have an impact on ASpR performance. Indeed, we observed a decline in ASpR performance when utilizing the activation vectors. However, this was expected since we concentrated the flow of information during the DNN training into some neurons only. This compact representation of speaker information may be the cause of some loss in information during the training process.

Despite the simplicity of this solution and as a first attempt, we believe that these results are encouraging. It is important to recall that our goal in this work is not to present a more accurate speaker recognition system, but rather to present a more explainable and a more easily interpretable ASpR model. Although not enhancing performance, the incorporation of activation function can be seen as a first step towards preparing the DNN model for explanations.

Nevertheless, it is also worth mentioning that the choices we made so far can be criticized. We are aware that this extractor is not without its limitations and that there is large room for further improvements. Indeed, the ResNet used in this extractor is not the most accurate ASpR system compared to recently proposed models such as WavLM and ECAPA-TDNN models. Given that the incorporation of thresholding function is indeed independent of the DNN architecture, these newer architectures may provide better performance, following the same process. However, still the challenges about these architectures explainability remains complex and unresolved. Furthermore, we are aware that the concept of binarizing vectors post-training may not be ideal. Instead, a more favorable approach would involve training a dedicated binary extractor designed to directly extract binary vectors. In an effort to explore a novel perspective for the BA-extractor, we propose in chapter 10 a different modelling of the attribute-based representations. This model extracts directly binary vectors and achieves improved performance while respecting our representation assumptions.

In the upcoming steps of this work, we adopt the BA-vectors generated from the Softplus-vectors as the binary-attribute-based speaker embeddings.

Chapter 7

Step 2: Binary-Attribute-based likelihood ratio estimation

7.1 Introduction . . . 86
7.2 The core concept of BA-LR . . . 86
7.3 Estimation of behavioral parameters . . . 89
7.3.1 Typicality . . . 90
7.3.2 Drop-out . . . 91
7.3.3 Drop-in . . . 92
7.4 Estimation of likelihood ratios . . . 92
7.4.1 Attribute LR estimation . . . 93
7.4.2 Global LR estimation . . . 96
7.5 Analyses and evaluation of speaker recognition performance . 97
7.5.1 Data sets and protocols . . . 97
7.5.2 Analyses of behavioral parameters . . . 98
7.5.3 Speaker recognition performance and generalization ability 100
7.6 Explainability of the LR . . . 100
7.6.1 Distribution of attribute-LLR values . . . 101
7.6.2 Shapley-like explanations . . . 102
7.7 Discussion and perspectives . . . 104

In this chapter, we build on the binary-attribute-based speaker embeddings mentioned earlier, for the second step of our approach. This step is designed to establish an interpretable and explainable framework to estimate the LR for speaker recognition. More specifically, we position ourselves in a forensic context where the LR is employed to evaluate the evidence. Even though we emphasize this context interpretability is a paramount, it is noteworthy that this framework might be applied more broadly to any ASpR system.

7.1 Introduction

ASpR systems are designed to determine whether two voice recordings belong to the same speaker using a machine that outputs a score. In ASpR literature, the predominant focus lies on enhancing performance, often neglecting the interpretability and explainability of the information which drives the output prediction. In a forensic context, the output of FASpR system is typically expressed as a LR, representing the ratio between two likelihoods corresponding to two competing hypotheses: either the two voice samples are spoken by the same person, or each voice sample is spoken by a different person. Even though the LR is meaningful and self-sufficient by nature [99], presenting a single number as the output of forensic automatic system is becoming a serious weakness with respect to regulatory compliance and ethical considerations [145, 135, 136]. Assessing the evidence necessitates an examination of the elements and factors that contribute to its value to be informed of any sort of discrimination bias and to help the court in their decision-making.

To address the aforementioned concern, this chapter introduces an interpretable and explainable framework for estimating the LR value, named Binary-attribute-based LR (BA-LR). In the following sections, we begin by presenting the core idea and motivations of BA-LR framework. Subsequently, an estimation of key parameters is proposed, followed by a LR estimation using these parameters. The results of applying BA-LR scoring scheme to an ASpR task are therefore presented, evaluating both performance and explainability aspects. Finally, we summarize and report conclusions and perspectives derived from this framework.

7.2 The core concept of BA-LR

In this section, we begin by highlighting the inherent lack of interpretability in LR calculation within a forensic ASpR system. Following that, we clarify the motivation behind introducing the BA-LR framework and draw inspiration from DNA individualisation process. Subsequently, we present our visionary BA-LR framework through an idealized example. Lastly, we outline some research questions that will be explored in this chapter.

Lack of interpretability in LR computation

Figure 7.1 introduces the lack of interpretability in the LR calculation in a scenario reporting the value of a vocal evidence.

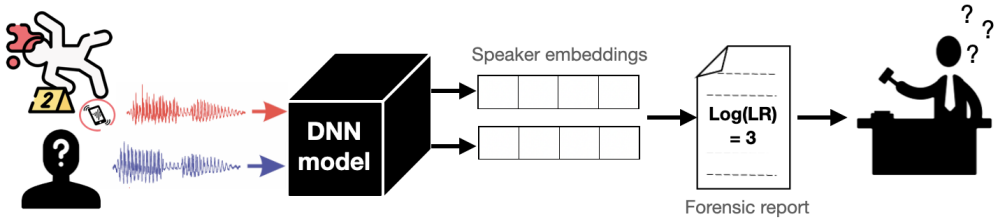


Figure 7.1: An illustration of the assessment of the value of speech evidence using ASpR system

Let us imagine a crime scene where a vocal trace, potentially containing the voice of the perpetrator (i.e. criminal), is retained as evidence. In this case, the vocal trace and a voice recording of a suspect speaker are both processed using a DNN-based ASpR system to extract corresponding speaker embeddings. Subsequently, a similarity score is computed and transformed into a LR value, which is then presented in court. Assigning a LR value of 20 (i.e. LLR=Log(LR)=3) effectively conveys that the evidence supports the prosecution's hypothesis 20 times more than the defense's hypothesis. According to the courts positions discussed in §3.4.1, our focus here lies in the case where the court requires further explanations about the value of evidence. In such case, simply relying on the LR value falls short. The court would inquire about the elements that influence the LR estimate and their properties to verify that the decision is not based on a discrimination bias [135, 136, 145, 147].

Motivation

Returning to the earlier illustrative example, Figure 7.2 shows how improvements can be made to offer a more interpretable perspective, moving beyond a mere support for one hypothesis over the other. Saying that the LLR equal to 3 is obtained from a composition of four LLRs, as shown in Figure 7.2, each dedicated to one factor and worth 2, -0.5, 2.5 and -1 respectively, not only indicates the same support for the prosecution hypothesis but it also provides a detailed composition of this support. In this context, one may say that this support comes mainly from the factors with $LLR = 2$ and $LLR = 2.5$. However, adding more information to the identified factors may grandly help the judge, jointly with the other case information, to take in hands the value of evidence. For instance, supplementing information about the intrinsic characteristics of the factors, such as their discriminant power and the expected reliability of their estimation, would offer more insights into the final LLR. The work in [214] emphasized the importance of involving the uncertainty of estimation (i.e. the probabilities of errors) related to LR estimation by saying that: *"With estimation comes always a degree of uncertainty, and it should be acknowledged that this uncertainty induces a risk of jumping to the wrong conclusions at court [...] Nevertheless, having the uncertainty in mind when reporting the likelihood ratio will reduce the risk of making an erroneous conclusion at court"*. Therefore, returning to our example, a LLR of 2.5 could arise from highly discriminant factor but with a low estimation reliability, while the LLR of -0.5 may be linked to a factor with a moderate discriminant power but with a very reliable estimation as illustrated in Figure 7.2.

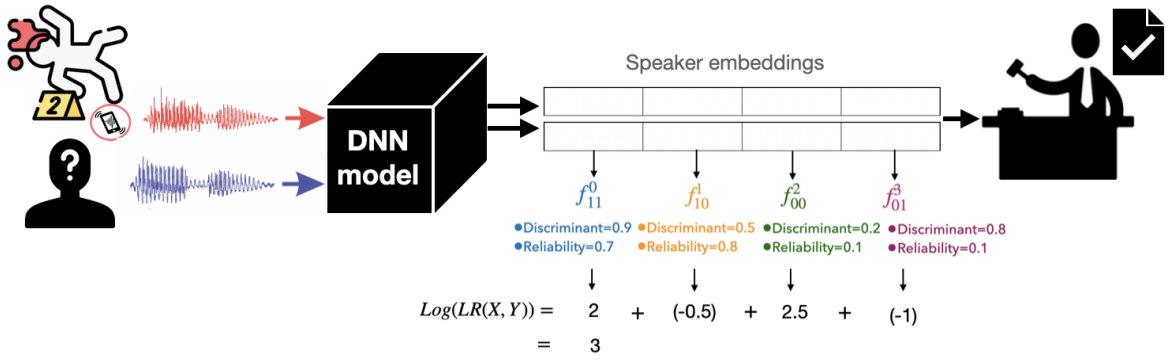


Figure 7.2: An illustration of an interpretable assessment of the value of speech evidence using ASpR system

This process is inspired from evaluating the evidence in forensic DNA individualisation, where the decomposition of the LR into partial LRs was firstly introduced [213]. This is further detailed in §5.2; To summarize the process, the expert compares a trace extracted from a crime scene and a suspect profile, both represented by a finite set of allele pairs at pre-defined loci [211, 212]. For each locus, the presence or absence of alleles in both parts of the comparison is used to estimate a partial LLR. Due to some sources of ambiguity [222], uncertainty by locus exists [214] and is quantified by drop-out and drop-in probabilities of alleles [218, 219]. These probabilities are also involved into the partial LLRs computations [216]. The global LLR is therefore obtained as the sum of the partial LLRs thanks to the independence between the loci involved.

The dreamlike BA-LR framework

Our novel BA-LR approach is primarily motivated by this process. The main goal is to present an interpretable computation of the final score as described by the previous example. In Figure 7.3, we describe the ultimate goal of our approach with a dreamlike toy scenario that illustrates our vision for the interpretable computation of LR.

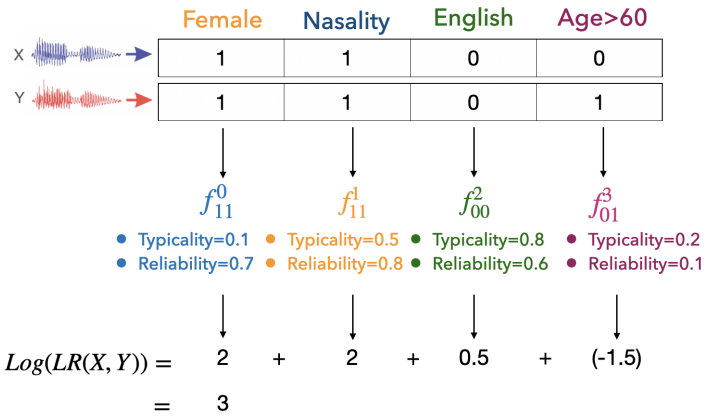


Figure 7.3: A dreamlike interpretable likelihood ratio calculation

Given two speech extracts X and Y, two binary-attribute-based embeddings are extracted, as demonstrated in chapter 6. Each dimension in these embeddings is repre-

sented by a voice attribute in a binary fashion, where a value of 1 denotes the presence of the voice attribute, while a value of 0 signifies its absence. Let us consider in this example a set of only four predefined voice attributes as shown in Figure 7.3. For each attribute k, a LLR is calculated denoted as f_{ij}^{k} taking into account both the values ij in both sides of the comparison (i.e. 00, 11, 01 or 10) and two behaviors related to the attribute that we model by typicality and reliability. Typicality [241] reflects how typical is the attribute among some population, while reliability indicates the degree of confidence in that attribute. Thus for this example, a potential interpretation of the final LLR value of 3 could be the following; the LLR of 3 means at the first place that the evidence supports, in the LR domain, the prosecution hypothesis 20 times more than the defence hypothesis. Additionally, this support is deriving from the contribution of four voice attributes including female voice, voice nasality, use of English language, and having an age of over 60 years. The most reliable voice attributes that are presenting the biggest contribution are mainly Female sex and voice nasality, with the former is rare attribute (i.e. highly discriminant) and the latter possessing a moderate typicality. It is thus becoming clear that the final value is driven by the contribution of discriminant attributes that are present in both samples. English language attribute is also reliable but very typical (i.e. not discriminant) among our population. Therefore, its absence in both comparison sides is adding a small contribution to the final value. We believe that providing this level of explanations when a LR is proposed can make it easier for decision makers to understand the existing information in the evidence.

Research questions

To bring this dreamlike framework to fruition, we aim in the next sections to address the following research questions:

- **RQ1**: How to estimate the behavior of a given attribute?
- **RQ2**: How to elaborate an interpretable formulation of the LLRs per attribute as a function of the attribute behavior?
- **RQ3**: Is the BA-LR framework applicable in a speaker recognition task?
- **RQ4**: Are we able to offer additional explanations for the final score? Which attribute behavior has a greater impact on the final score?

7.3 Estimation of behavioral parameters

In BA-LR approach, each voice attribute, denoted as BA, is characterized by two key behaviors: its typicality and its reliability. Drawing inspiration from the DNA individualisation process, as detailed in §5.2, these behaviors are represented in our approach through three behavioral parameters per BA; The first parameter is the typicality [242] that reflects the attribute's discriminant power. The two remaining parameters concern the reliability that quantifies two probabilities of errors associated with the attribute. These probabilities were initially incorporated in the computation of LR in forensic DNA analysis, namely the Drop-in and Drop-out [220, 216, 222].

In this section, we present the definitions of the behavioral parameters. Subsequently, we introduce a simple and a straightforward proposal for their estimation, inspired by DNA concepts. **Please note that these estimations are may be not the best, they are chosen to mainly prioritize simplicity and ease of understanding**.

7.3.1 Typicality

We define the typicality of an attribute BA_i, denoted as T_i, as the frequency of speaker pairs sharing the attribute in the relevant population. The typicality reflects the discriminative power of an attribute [243, 242, 216]. The more typical an attribute is, the more frequent it is in the reference population, the less discriminating it is [243]. We propose a basic estimation of the typicality in Equation (7.1) calculated as the number of speaker couples sharing that attribute divided by the total number of couples in the reference population.

$$T_i = \frac{\sum^{N_c} \mathrm{P}_{S_1} \cap \mathrm{P}_{S_2} = \{\mathrm{BA}_i = 1\}}{N_c} \tag{7.1}$$

Where P_{S_1} and P_{S_2} are the profiles of the two speakers S_1 and S_2 in the couple and N_c is the number of speaker couples in the relevant population. The notions of *speaker profile* and *relevant population* are clarified and further explained in the following.

The relevant population

The relevant population is defined by [101, 100, 242] as the set of speakers chosen when formulating the defence hypothesis in forensic speaker recognition. Drygajlo et. al in [100] state that: "Although possible in theory that the different-speaker hypothesis includes any possible speaker, this hypothesis is usually more restricted at least to speakers of the same sex, but often also to speakers of the same language and perhaps even the same language variety". They also state that the definition of the relevant population could be related to technical procedure such as an ASpR model that is already trained on a set of speakers, or it could be defined based on an authority request of specific profiles related to the case. *"If the mandating authority or party requests a specific relevant population, it needs to be assessed by the forensic expert whether this request can be met. If the mandating authority or party does not request any specific relevant population it is necessary that the forensic expert defines a relevant population that is compatible with the circumstances of the case at hand and for which the necessary databases and other resources are available."*[100].

In this work, our relevant population comprises the speakers within the training data set of our DNN model. This is because the attributes are learned based on this population, and also because a larger population allows us to cover more cases and gain a wider view on the behavior of these attributes.

The "elusive" speaker profile

As discussed in §5.2, unlike the DNA profile that comprises all information about an individual, **the notion of a speaker voice profile is impractical and close to unachievable**. This is due to the various variabilities inherent in our voice, often beyond control, in contrast to the more stable nature of DNA. Additionally, a crucial factor is the linguistic content, bringing forth additional variabilities. The content is consistently changing. Due to brief speech segments or the infrequent occurrence of certain phonemes by the speaker, this aspect would highly impact the definition of the speaker profile. This gives raise to the drop-out phenomenon defined later in this section.

Thus, in this work, we set our own definition of the speaker profile. For a given speaker S_j, we consider an attribute, BA_i, present in his profile P_{S_j}, when it is present in at least one speech extract or utterance U of the speaker as expressed in Equation (7.2).

$$P_{S_j}(\mathrm{BA}_i) = \begin{cases} 1 & \text{if } \sum_{U \in S_j} U(\mathrm{BA}_i) \geq 1 \\ 0 & \text{Else} \end{cases} \tag{7.2}$$

While aware that this definition may not align with reality, we are adopting a generous position toward the speaker profile, minimizing the risk associated with the potential presence of the attribute. In our view, it is safer and more prudent to regard the attribute as present if it occurs at least once, rather than searching for an optimal presence threshold for each attribute, which is never evident. This choice is reinforced by the incorporation of the drop-out notion, which re-compensates any potential overestimation of the attribute's presence.

7.3.2 Drop-out

The drop-out is firstly defined in DNA field [222, 216, 221, 218, 217] as the disappearance of some alleles from the profile due to technical process (refer to §5.2). It might occur in two scenarios: 1) a *false negative detection* of the attribute. 2) a non-presence of the attribute due to speech variability or due t insufficiency of data. Here, we define it as the probability that an attribute does not appear in a given utterance while it was observed in at least one other utterance of the considered speaker. The drop-out is directly linked to the fact that a given speech extract is not containing all the speech information about a given speaker.

In this work, drop-out probability of an attribute is estimated firstly per each speaker, denoted by Dout_i^S, as expressed in equation (7.3). It is calculated as the number of utterances U^S, not having a BA_i given that this BA_i is present in the profile of speaker S, divided by the number of utterances of that speaker N_S. Then a drop-out of an attribute, denoted as $Dout_i$, is averaged over all the N speakers having that attribute in the relevant population as expressed in equation (7.4). This two steps

estimation is followed because the number of utterances is variable amon speakers.

$$\text{Dout}_i^S = \frac{\sum_{U \in S}^{N_S} (U(\text{BA}_i = 0) | P_S(\text{BA}_i) = 1)}{N_S} \tag{7.3}$$

$$\text{Dout}_i = \frac{\sum_{j=1}^{N} \text{Dout}_i^{S_j}}{N} \tag{7.4}$$

7.3.3 Drop-in

Similarly to attribute drop-out, drop-in phenomenon is directly inspired from DNA process [220, 222]. It might occur in speech due to noise caused by multiple factors such as environment, recording conditions, quality of device...etc. In this work, we define drop-in as the probability of encountering a noise leading to the false presence of an attribute in a given speech utterance. It can be considered as a *false positive* detection of the attribute. The Drop-in phenomenon is considered as independent of attributes as it reflects the general noise in the data [222]. It is estimated with a noise factor *Din* multiplied by T_i, the typicality of the attribute, expressed in Equation (7.5). This represents the idea that a drop-in occurred in an attribute along with its presence frequency T_i.

$$P_{din} = Din * T_i \tag{7.5}$$

Even in DNA, it was shown to be difficult to estimate the drop-in. Gill et.al in [216] say that:

> *"There is no absolute method to determine if drop-in or contamination has occurred in a casework sample, but negative controls can be used to estimate the probability of drop-in within casework samples" [...] "To calculate the risk of a drop-in event in a casework profile, we multiply together the probability of drop-in with the probability of the specific allele a that is conditioned to have dropped in: Pr(C)pa [220]".*

7.4 Estimation of likelihood ratios

The goal of the proposed BA-LR method is to decompose the LR into attribute LRs and subsequently employ an interpretable formulation to estimate these attribute LRs, incorporating the attribute behavioral parameters [13]. Figure 7.4 presents an overview of BA-LR approach; Given two speech samples X and Y representing a suspect recording and a vocal trace, respectively, we firstly extract the corresponding BA-vectors using step 1, as outlined in Chapter 6. Subsequently, we calculate an LR for each attribute. This calculation takes into account the value of the attribute on both sides of the comparison, along with the attribute's behavioral parameters estimated from the training dataset. The global LR value assessing the comparison pair is therefore computed as the product of the attribute LRs.

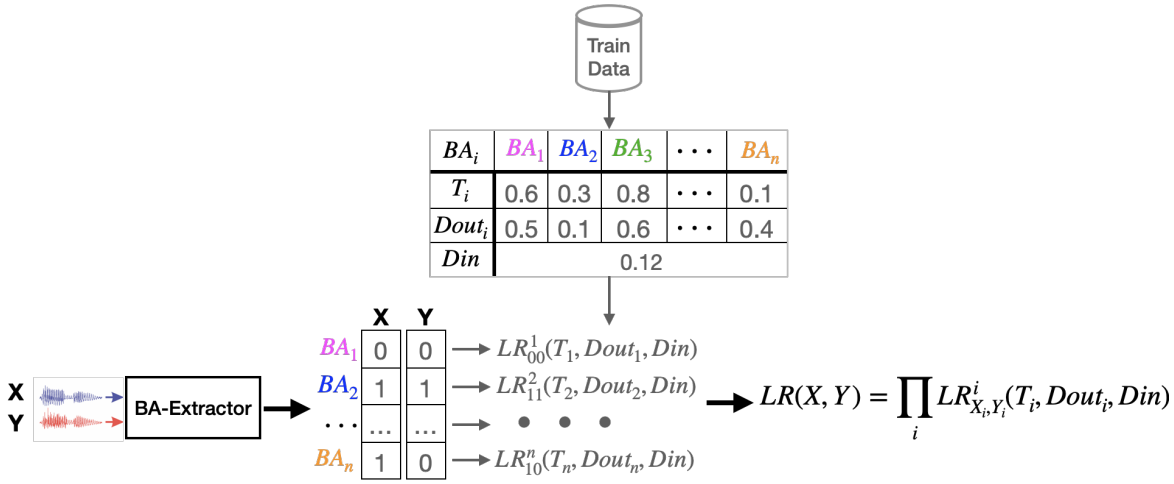


Figure 7.4: Global likelihood ratio estimation using BA-LR

In this section, we introduce two versions for attribute LRs estimation, along with their interpretations within a forensic context. Subsequently, we infer the final LR value calculated using these attribute LRs for a pairwise comparison.

7.4.1 Attribute LR estimation

The estimation of the attribute LRs is inspired from [216, 222, 223, 221]. An attribute LR is computed following two hypotheses, expressed in Equation (7.6), as the ratio of the probability of the attribute given prosecution hypothesis H_p divided by the probability of the attribute given the defence hypothesis H_d.

$$LR^i_{X_i,Y_i} = \frac{P(X_i, Y_i|H_p)}{P(X_i, Y_i|H_d)} \tag{7.6}$$

Based on binary values of BA_i, four potential cases of (X_i, Y_i) could be encountered and therefore four values of LRs per attribute i are considered such as $LR^i_{0,1}$, $LR^i_{1,0}$, $LR^i_{0,0}$ and $LR^i_{1,1}$. A first case is when the attribute is present in X and not in Y (e.g. BA_n in Figure 7.4). A second case is when it is absent in X but present in Y. A third case is when the attribute is absent in both X and Y (e.g. BA_1 in Figure 7.4). The forth case is when the attribute is present in both X and Y (e.g. BA_2 in Figure 7.4).

Based on the four cases just described, we propose two versions of attribute-LRs estimation. The first version is referred to as **DNA-inspired Attribute LR estimation**, while the second version is named **Speech-adapted Attribute LR estimation**. The difference between both versions is mainly driven by the assumptions set in each version. In the following, both proposals are presented.

DNA-inspired

This version is referred to as a *naive version*, wherein we apply the same reasoning as done in DNA individualisation [222], adopting the following assumptions:

- The drop-in and drop-out phenomena could occur **only** in the vocal trace.
- The suspect recording is considered as the reference of the suspect's voice.
- The case in which a drop-in could occur is reflected by a factor Din multiplied by T_i, the typicality of the attribute, which means that a drop-in happened in BA_i with probability T_i as done in [222]. The case in which there is no drop-in, it is irrelevant to multiply by the complementary of the typicality.

We describe in the following, the forensic hypothetical rationale for formulating the attribute LRs in the four cases, expressed in Equation (7.7):

- *Y: (BA_i=0) , X: (BA_i=1)*: **Under $\mathbf{H}_p$**, the numerator, the prosecutor asserts that the drop-out observed at BA_i in the trace is simply a result of inherent variability. However, in reality, the trace is attributed to the suspect. **Under $\mathbf{H}_d$** hypothesis, in the denominator, the defense explores all potential scenarios involving random individuals. Consequently, the defense argues that the drop-out at BA_i in the trace may be a consequence of an attribute-specific error. Alternatively, it could be linked to another speaker in the population for whom BA_i was not observed, eliminating the possibility of a drop-out.

- *Y: (BA_i=0) , X: (BA_i=0)*: **Under $\mathbf{H}_p$** there is a 100% correspondence at BAi, the trace is certainly belonging to the suspect. Conversely, **under $\mathbf{H}_d$**, the defense acknowledges the absence of BAi in the trace and additionally proposes that the BA_i might be present in the trace, but a drop-out could have occurred.

- *Y: (BA_i=1) , X: (BA_i=1)*: **Under $\mathbf{H}_p$**, there is a 100% correspondence at BAi, it is absent in both samples. Conversely, **under $\mathbf{H}_d$**, the defence would argue that there might be no drop-out in the trace, or perhaps BAi is absent in the trace, but a drop-in caused its appearance.

- *Y: (BA_i=1) , X: (BA_i=0)*: **Under $\mathbf{H}_p$**, a drop-in occurred, leading to the appearance of BAi in the trace, but in fact the trace belong to the suspect. However, **under $\mathbf{H}_d$**, the defence is faced to two scenarios: there might have been a drop-in, or perhaps there was no drop-in, and alternatively, the trace could belong to someone else in the population.

$$LR^i_{X_i,Y_i} = \begin{cases} \dfrac{\text{Dout}_i}{T_i \cdot (\text{Dout}_i + \overline{\text{Dout}_i})} & \text{if Y}(\text{BA}_i = 0), \text{X}(\text{BA}_i = 1) \\ \dfrac{1}{T_i \cdot (\overline{\text{Din}} + \text{Dout}_i)} & \text{if Y}(\text{BA}_i = 0), \text{X}(\text{BA}_i = 0) \\ \dfrac{1}{T_i \cdot (\overline{\text{Dout}_i} + \text{Din} \cdot T_i)} & \text{if Y}(\text{BA}_i = 1), \text{X}(\text{BA}_i = 1) \\ \dfrac{\text{Din} \cdot T_i}{T_i \cdot (\text{Din} \cdot T_i + \overline{\text{Din}})} & \text{if Y}(\text{BA}_i = 1), \text{X}(\text{BA}_i = 0) \end{cases} \tag{7.7}$$

Equation (7.7) illustrates the proposed attribute LRs computation for the four cases, where $\overline{Din} = 1 - Din$ and $\overline{Dout} = 1 - Dout$. These formula are elaborated upon prosecution and defence hypotheses based on Equation (7.6).

This formulation is mainly built on the assumption that the phenomena of drop-in and drop-out might occur only in the trace sample. However, unlike DNA where we have always a ground truth about the suspect, in a speech comparison case, errors could occur in both sides of the pairwise comparison. As a first attempt, we opt to merge the cases 01 and 10 into one case, by symmetrizing them as expressed in Equation (7.8).

$$LR^i_{0,1} = LR^i_{1,0} = \frac{LR^i_{0,1} + LR^i_{1,0}}{2} \tag{7.8}$$

Indeed, this solution unifies the different case attribute-LR but it may not be entirely logical. A more proper solution is proposed in the second version where the case 01 and 10 are initially formulated similarly.

Speech-adapted

This version is referred to as a *more logical* version. In this version, different assumptions are made as follows:

- Drop-out and drop-in phenomena could occur in both, X and Y recordings.
- Drop-out and drop-in phenomena occurring in X are independent of those occurring in Y.
- Similarly to DNA-inspired, the absence drop-in is represented by $\overline{Din}$, whereas the absence of drop-out is reflected by $\overline{Dout_i}$.
- (X_i, Y_i) is reflected by two states. The observed state in the time of comparison, and the actual state without the impact any misleading phenomenon.

The forensic hypothetical rationale for formulating the attribute LRs in the four cases, for this version, is described in the following. Please refer to Appendix.B for a visual representation of all combinations:

- *Y: (BA_i=0) , X: (BA_i=0)*: The observed state is $(0, 0)$. **Under** H_p, the prosecution considers two possibilities: either the true state is also $(0, 0)$, or it is

$(1,1)$ but with drop-out on both sides resulting in $(0,0)$. **Under** H_d, the defense presents various scenarios. He argues that the true state could be either $(0,1)$ or $(1,0)$, but with drop-out on one side, leading to $(0,0)$. Thus, this possibility is counted twice. Moreover, it is possible that with drop-out on both sides, $(0,0)$ is observed from the true state $(1,1)$. Additionally, if there are no drop-ins on either side, $(0,0)$ is obtained from the true state $(0,0)$.

- *Y: (BA_i=1) , X: (BA_i=1)*: **Under** H_p, there is a 100% correspondence. Alternatively, if the true state is $(0,0)$ but experiences drop-out on both sides, this would lead to observe $(1,1)$. **Under** H_d, the true state could be either $(0,1)$ or $(1,0)$, but with a drop-in on one side or the other, resulting in $(1,1)$. Additionally, it is possible that with a drop-in on both sides, we observe $(1,1)$ from the true state $(0,0)$. Furthermore, there may be no drop-outs on either side.

- *Y: (BA_i=1|0) , X: (BA_i=0|1)*: **Under** H_p, the observed state is $(1,0)$ or $(0,1)$, but they should belong to the same speaker. Thus, it is possible that the true state was $(0,0)$, experiencing a drop-in on one side but not the other. Alternatively, the true state could be $(1,1)$, but a drop-out occurred on only one side, not the other. **Under** H_d, since both samples belong to different speakers, it is conceivable that the true state is $(0,1)$ or $(1,0)$. There could be a drop-in on one side and not the other. A dropout on one side but not the other. A simultaneous drop-in on one side and a dropout on the other side.

Equation system (7.9) presents the mathematical formulation of this reasoning following the forensic hypothesis.

$$LR^i_{X_i,Y_i} = \begin{cases} \dfrac{1+\text{Dout}_i^2}{T_i \cdot (2 \cdot \text{Dout}_i \cdot \overline{\text{Din}} + \text{Dout}_i^2 + \overline{\text{Din}}^2)} & \text{if}(\text{BA}_i^Y = 0, \text{BA}_i^X = 0) \\ \dfrac{1+(\text{Din} \cdot T_i)^2}{T_i \cdot (2 \cdot \text{Din} \cdot T_i \cdot \overline{\text{Dout}_i} + (\text{Din} \cdot T_i)^2 + \overline{\text{Dout}_i}^2)} & \text{if}(\text{BA}_i^Y = 1, \text{BA}_i^X = 1) \\ \dfrac{\overline{\text{Din}} \cdot \text{Din} \cdot T_i + \text{Dout}_i \cdot \overline{\text{Dout}_i}}{T_i \cdot (\overline{\text{Din}} \cdot \text{Din} \cdot T_i + \text{Dout}_i \cdot \overline{\text{Dout}_i} + 1 + \text{Din} \cdot T_i \cdot \text{Dout}_i)} & \text{if}(\text{BA}_i^Y = 0, \text{BA}_i^X = 1) \\ \dfrac{\overline{\text{Din}} \cdot \text{Din} \cdot T_i + \text{Dout}_i \cdot \overline{\text{Dout}_i}}{T_i \cdot (\overline{\text{Din}} \cdot \text{Din} \cdot T_i + \text{Dout}_i \cdot \overline{\text{Dout}_i} + 1 + \text{Din} \cdot T_i \cdot \text{Dout}_i)} & \text{if}(\text{BA}_i^Y = 1, \text{BA}_i^X = 0) \end{cases} \tag{7.9}$$

This formulation is shown to be more logically presented and interpreted, since it considers a real voice comparison case. Here, based on the followed reasoning, the case 01 and 10 are automatically behaving similarly.

7.4.2 Global LR estimation

To be able to provide a global LR as the product of the attribute LRs (see Figure 7.4), the assumption of independence between attributes should be respected in both versions. Referring to the previous chapter 6, we demonstrated experimentally, in

§6.4.3, that the attributes in the BA-vectors fulfill this assumption. As a result, the global LR is calculated as shown in Equation (7.10):

$$LR(X,Y) = \prod_{i=1}^{n} LR^{i}_{X_i,Y_i} \tag{7.10}$$

In practice, the logarithmic versions of the LRs are used in order to move into the Log(LR) domain and take advantage of its additive nature. In light of this, the Log of the final LR becomes the sum of the attribute LLRs as shown in Equation (7.11).

$$LLR(X,Y) = \sum_{i=1}^{n} LLR^{i}_{X_i,Y_i} \tag{7.11}$$

7.5 Analyses and evaluation of speaker recognition performance

In this section, our goal is to evaluate the overall performance of BA-LR approach in terms of speaker discrimination and generalization capabilities across different datasets. To achieve this, we begin by outlining the datasets employed and their corresponding evaluation protocols. Subsequently, we conduct some analysis of the behavioral parameters followed by a comparative analysis of the speaker verification performance between the two versions of our approach and the baseline x-vector system.

7.5.1 Data sets and protocols

Table 7.1: Description of data sets

	Train	**Test**		
	VoxCeleb2	**VoxCeleb1**	**SITW**	**VOiCES**
# of speakers	5,994	1,251	180	100
# of utterances	1,021,175	153,516	2,883	11,392
# of comparison pairs	35, 964[2]*2[1]	56,279*2[1]	3,658*2[1]	36,443*2[1]

[1] 1 for target and 1 for non-target comparison pairs

[2] The number of target pairs is composed from all speakers with only 3 utterances for each speaker.

In this experiment, we use four corpora such as: VoxCeleb1&2, SITW [244] and VOiCES [245], as summarized in Table 7.1. During steps 1 and 2 of our BA-LR approach, VoxCeleb2 data set is used for training the BA-extractor and computing behavioral parameters related to attributes (T_i, $Dout_i$, Din). VoxCeleb1, SITW, and VOiCES are used for testing only, and have no intersection with VoxCeleb2 in terms of speakers. While VoxCeleb1&2 are extensively detailed in §6.4.1, the descriptions of SITW and VOICES are provided below.

- **SITW**[1]. It is composed of speech samples in English from open-source media representing unconstrained or wild conditions and pronounced by 180 speakers. We use the evaluation corpus of SITW dataset [244] that gives 3,658 target pairs and we select randomly the same number of non-target pairs.
- **VOiCES**[2]. It is composed of excerpts pronounced by 100 speakers in acoustically challenging and reverberant environments. We use the evaluation of "VOiCES Distance Speaker Recognition Challenge 2019" [245, 246]. It included 36,443 target pairs and we select randomly 36,443 non-target pairs.

7.5.2 Analyses of behavioral parameters

To experimentally estimate the attribute behavioral parameters, it is necessary to first define our relevant population. The training dataset, VoxCeleb2, consisting of approximately 6000 speakers, is employed here without any data augmentation to accurately represent our relevant population. All the parameters, T_i, $Dout_i$ and Din factor, are calculated using VoxCeleb2.

To have an overview about the speakers profiles in terms of the number of attributes, Figure 7.5 presents the distribution of present attributes in speakers profiles of the training data. It is important to recall that the initially extracted number of attributes on VoxCeleb2 is 256 per BA-vector without considering dead neurons (refer to chapter 6). As can be seen in the figure, the majority of speakers exhibit between 150 and 175 present attributes out of 256. Towards the extremes of the distribution, there are a few speakers with certain present attributes (75-100) and others with over 200 attributes.

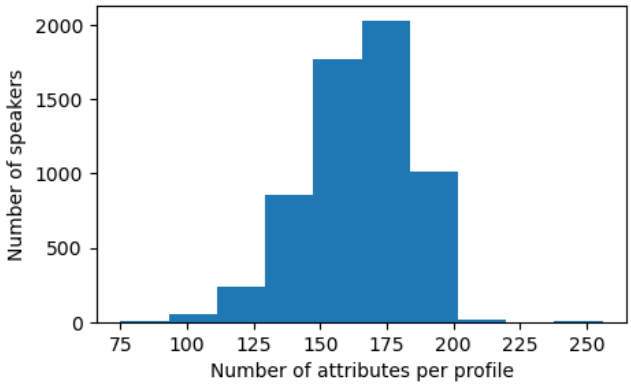


Figure 7.5: The number of speakers as a function of the number of present attributes in their profiles

Based on these profiles, typicality and drop-out parameters are estimated. Figure 7.6 illustrates the distributions of attributes typicality and drop-out values. Approximately 50% of the BAs exhibit a typicality in the range of $[0.6, 0.8]$, indicating a

[1]http://www.speech.sri.com/projects/sitw/
[2]https://iqtlabs.github.io/voices/

modest discrimination power, while around 25% display a high discrimination power with a typicality of approximately $[0.15, 0.57]$. The drop-out values range from 0.4 to 0.8. This high range could be attributed to short speech extracts that may not capture all speech and linguistic variabilities or variations in the number of utterances among speakers. The high values of typicality and drop-out parameters, demonstrated in the distributions, are indeed a direct consequence of their reliance on the speaker profile. Also, they could be seen as indicative of the quality of BA-vectors in the training data.

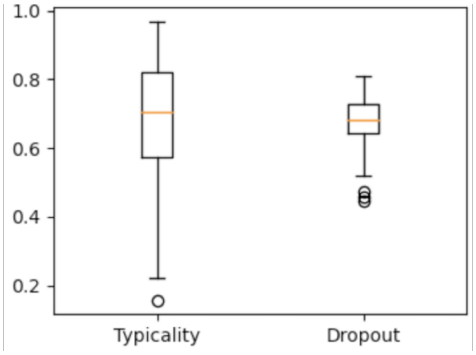
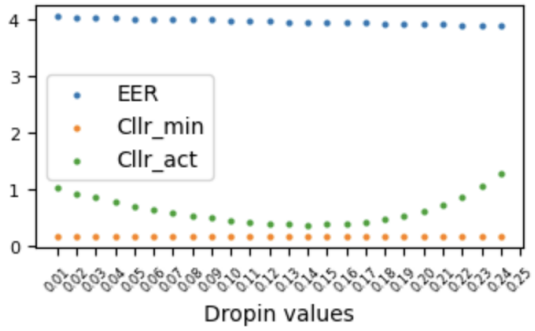
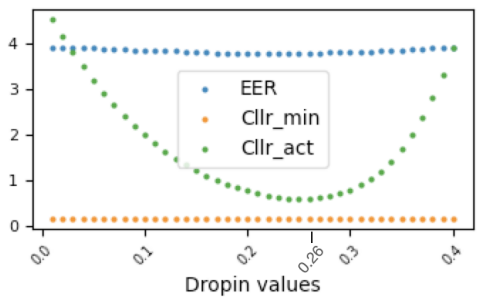


Figure 7.6: Distributions of attributes typicality and drop-out values

Drop-in probability, on the other hand, is estimated based on the *Din* parameter. This parameter is tuned and optimised using a set of comparison pairs from the train dataset (refer to Table 7.1). The evaluation process is performed using both versions of BA-LR estimation to estimate a *Din* value for each version. Given a set of *Din* values, the optimisation of this parameter is performed in such a way that the C_{llract} [127] and the C_{llrmin} (refer to 3.3.3) of the train comparison pairs become as close as possible. This process is described in Figure.7.7 under both versions of BA-LR. The convergence is quite regular, with an optimum in a flat region around [0.11,0.15] for DNA-inspired (Figure.7.7a), giving an optimum value for *Din* of ∼0.12. For Speech-adapted (Figure.7.7b), the flat region is around [0.24,0.27], with an optimum *Din* value of ∼0.26.



(a) Using DNA-inspired, $Din = 0.12$



(b) Using Speech-adapted, $Din = 0.26$

Figure 7.7: Estimation of the optimal value of Din on the training data

7.5.3 Speaker recognition performance and generalization ability

Table 7.2: Speaker recognition performance in terms of EER and $C_{llrmin/act}$

Dataset	X-vectors		BA-vectors			
	Cosine		DNA-inspired BA-LR		Speech-adapted BA-LR	
	EER	$C_{llrmin/act}$	EER	$C_{llrmin/act}$	EER	$C_{llrmin/act}$
Vox1	1.37%	0.06/0.82	3.70%	0.14/0.31	3.50%	0.13/0.48
SITW	1.40%	0.06/0.82	3.50%	0.13/0.28	4.00%	0.14/0.49
VOiCES	3.96%	0.15/0.87	4.70%	0.18/0.46	5.12%	0.19/0.89

The speaker recognition performance of the BA-LR approach is assessed by EER and $C_{llrmin/act}$. Table 7.2 presents the performance of the two versions of BA-LR and X-vector baseline for speaker discrimination, evaluated on three datasets. The overall performance of BA-LR shows its good discriminative power across the three datasets, with respect to the well-known trade-off between performance and explainability. Despite having on average 1.7% and 1.96% absolute increase in EER, for DNA-inspired and speech-adapted versions respectively, compared to the X-vector system, we believe that this loss is rather small. This is to be compared with the important dimensionality reduction, from 8192 bits for X-vectors of 256 floats to 205 bits for BA-vectors. Additionally, one crucial aspect to mention is that BA-LR scoring retains only the information that it is able to explain. As can be observed, in terms of EER, speech-adapted is slightly better than DNA-inspired on Vox1, the closest set to the training set, and slightly worse for the other sets which far from train conditions. Despite the small difference between C_{llrmin} and C_{llrAct} for BA-LR, 0.2 for DNA-inspired and 0.46 for Speech-adapted, the LLRs are shown to be poorly calibrated. Considering this, a subsequent calibration step should be taken into consideration.

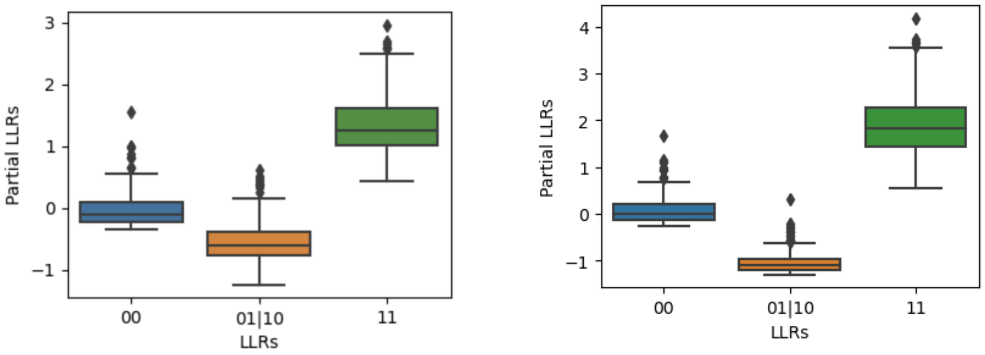
As this work primarily aims to demonstrate the interpretability and explainability of LR, in the next section we will solely focus on the VoxCeleb1 comparison pairs.

7.6 Explainability of the LR

This section aims to explain the contribution of each attribute-LLR to the final LLR. Before delving into the explanation of LLRs, it is relevant to first examine the distribution of the attribute LLRs, under the two versions of BA-LR. Subsequently, we draw an analogy to the post-hoc SHAP method (refer to §4.4.3) for estimating the contribution of each attribute to the final LR estimation.

7.6.1 Distribution of attribute-LLR values

Figure 7.8 illustrates the distribution of attribute-LLR values for the three cases: 00, 11, and 01|10[3], for both BA-LR versions, computed on VoxCeleb1 comparison pairs. As expected, the LLR_{11} values are predominantly positive for DNA-inspired and Speech-adapted, confirming the robust presence of an attribute on both sides of a pair. Conversely, the $LLR_{01|10}$ values are generally negative for DNA-inspired and Speech-adapted, except for some outliers, indicating a conflict when a given attribute is present in only one side of the pair. However, this distribution is more adjusted for Speech-adapted with only one outlier, as indicated in Figure.7.8b, than for DNA-inspired in Figure.7.8a. The Speech-adapted distribution for this case seems to be more logical than DNA-inspired. Lastly, LLR_{00} values are primarily centered around 0 for both versions, contributing minimally, with occasional positive high exceptions. This aligns with the notion that sharing the absence of a rare feature imparts little information, while sharing the absence of a highly present feature can significantly influence the final value.



(a) Using DNA-inspired BA-LR (b) Using Speech-adapted BA-LR

Figure 7.8: Distribution of attribute LLRs values for 00, 11 and 01|10 cases, computed on VoxCeleb1 comparison pairs

A more dedicated comparison between the attribute LLRs of both versions of BA-LR is demonstrated in Figure.7.9. This figure shows the relationship between attribute-LLRs of DNA-inspired and Speech-adapted. As can be observed, LLR_{11} in Speech-adapted are having positively higher values than DNA-inspired. LLR_{00} are shown to behave almost similarly in Speech-adapted and DNA-inspired with a very slight translation to a bit higher values. In contrary to DNA-inspired, $LLR_{01|10}$ values are shown to be all localized under zero for Speech-adapted, except for one outlier.

[3]The case "01" or "10"

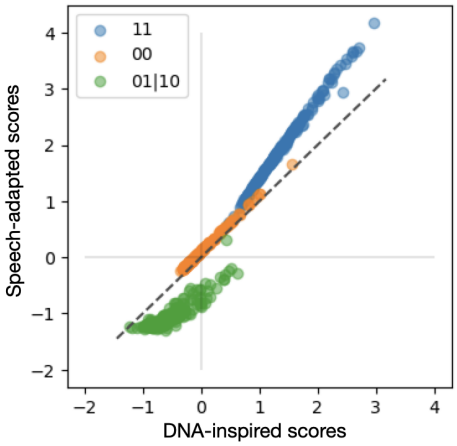


Figure 7.9: Relationship between attribute-LLRs of DNA-inspired and Speech-adapted versions

7.6.2 Shapley-like explanations

Referring to the Shapley values [197] properties described in §4.1, one may notice that the attribute LLRs possess these same properties, including Efficiency, Symmetry, Linearity, and dummy. This enables us to leverage the complete capabilities of SHAP functionalities[4] to quantify the impact of each speech attribute on the final LLR value (See Figure 4.6). From this perspective, the attribute LLRs are regarded as Shapley-like values, serving to quantify the contribution of each attribute to the final LLR. Using this approach, we explore explanations of the BA-LR framework, creating a transparent and user-friendly system for forensic practitioners, ensuring ease of interpretation in a court setting. Thus, we provide two types of explanations for the final LR: at the local level for a single comparison pair, and at the global level averaging all local explanations across all comparison pairs. This is further detailed in the remaining of this section.

Local explanations

The local explanations concern a single observation. These explanations are given by the contributions of attributes to the final LLR of this observation. It is important to note that we provide here an example of a local explanation using the DNA-inspired of BA-LR. This is to avoid repetition, since the same reasoning could be applied for each observation using either version.

Figure 7.10 shows the type of explanation SHAP provides for two single predictions, a target (above) and a non-target (below) comparison pairs. This figure is called a force plot [199]; it follows the basic idea that each input attribute contributes a force to push the model towards a certain output [183]. The forces, polygon widths, are the calculated Shapley-like values (i.e. attribute LLRs), and each prediction is considered

[4]https://github.com/slundberg/shap

starting from the base value. The value $f(x)$ represents the predicted value, and the forces are made to balance on the prediction, similar to balancing weights on a seesaw. The value 3 given to the attribute is a category indicating that the attribute is present in both sides of the comparison 11. Unless not shown in the figure, value 2 and 1 are given as categories to 00 and 01|10, respectively. In these examples, the BA-LR approach is correctly estimating the target and non-target pairs. The attributes in red push the prediction towards the positive direction, and the blue attributes push the prediction to the inverse direction. It is worth noting that the overall LLR value can be driven by a few LLRs with high positive value or by a majority of partial LLRs with negative values. According to the force plot, the most significant contributing factors toward the prediction are BA_9, BA_{223}, BA_{224} and BA_{110},BA_{25}, BA_{224} for target and non-target predictions, respectively. More information about these most contributing attributes are further detailed in Table 7.3.

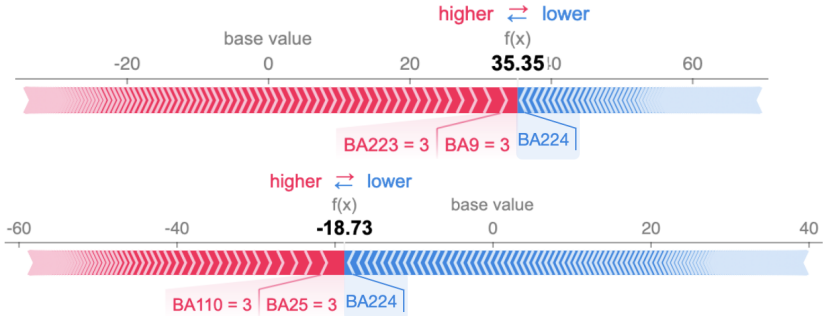


Figure 7.10: Attributes contributions to the final LLR for a target (above) and a non-target (below) comparison pairs.

Table 7.3: Details about the most contributing attributes for two speech pairs, a target pair and a non-target pair

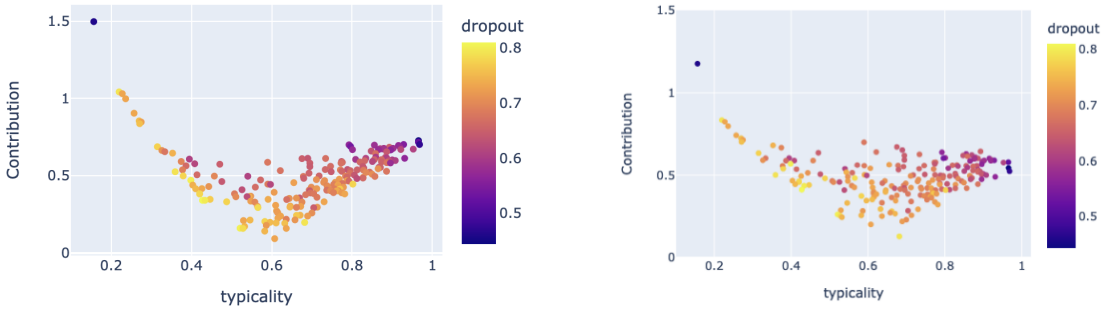
	target pair		**non target pair**		
	BA9	**BA223**	**BA110**	**BA25**	**BA224**
(X_i,Y_i)	(1,1)	(1,1)	(1,1)	(1,1)	(0,1)
Attribute LLR	2.43	2.32	2.0	2.96	-1.23
Typicality	0.15	0.39	0.37	0.21	0.96
Dropout	0.45	0.80	0.68	0.79	0.44
Final LLR	35.35		−18.73		

Examining Table 7.3 reveals a noteworthy observation: the presence of the most influential attributes, specifically denoted as $(1, 1)$ on both sides of the comparison, significantly contributes to the final LR with notably high values. These attributes exhibit lower typicality values compared to others, reflecting their discriminant power. Recognizing that the final LR value is influenced predominantly by these discriminant

attributes provides persuasive evidence, whether in the positive direction, as exemplified by BA_9, or in the negative direction, as illustrated by BA_{224}. As addressed earlier, the dropout values have been demonstrated to be generally high. However, given this Table, one may reasonably deduce that attributes BA_{223} and BA_{25} are most likely not reliable attributes because of the high probability of $Dout$.

Global explanations

Global explanations are provided by averaging all the local explanations across a set of comparison pairs. This gives a much better understanding of attribute importance. Here, we provide a global description for each version of BA-LR.



(a) Using BA-LR (DNA-inspired) (b) Using BA-LR (Speech-adapted)

Figure 7.11: Average contribution of attributes among all pairs Vs. their behavioral parameters, typicality & drop-out

Figure 7.11 elaborates a relationship between the attributes average contribution to LLR and their associated behavioral parameters, using the two versions. Each point in the figure represents one attribute. For DNA-inspired, clearly, attributes with typicality values ranging from 0.15 to 0.4, as well as those with values from 0.8 to 1, drive most of the contribution. As expected, the high contribution from the former is regarded as reasonable, considering their strong discriminatory capability. It is even straightforward and more easy to explain for the latter, given their low discriminant power but high reliability, indicated by the low dropout values. For Speech-adapted, we can observe a practically similar behavior but less clear. Apart from the extremities, the attributes presenting moderate typicality values ranging from 0.4 to 0.8 in DNA-inspired are weakly contributing to the output with contribution values below 0.5. This could be linked to the fact that their contribution does not decisively influence the output. This is also the case for Speech-adapted with some outliers.

7.7 Discussion and perspectives

In this chapter, we presented the second step of our approach, which builds upon the binary-attribute-based embeddings extracted in the first step in the previous chapter.

While existing speaker recognition solutions fall short of delivering an acceptable level of interpretability and explainability, our aim in this work is to introduce a novel solution that establishes an interpretable and explainable framework for calculating the LR in a speaker recognition task.

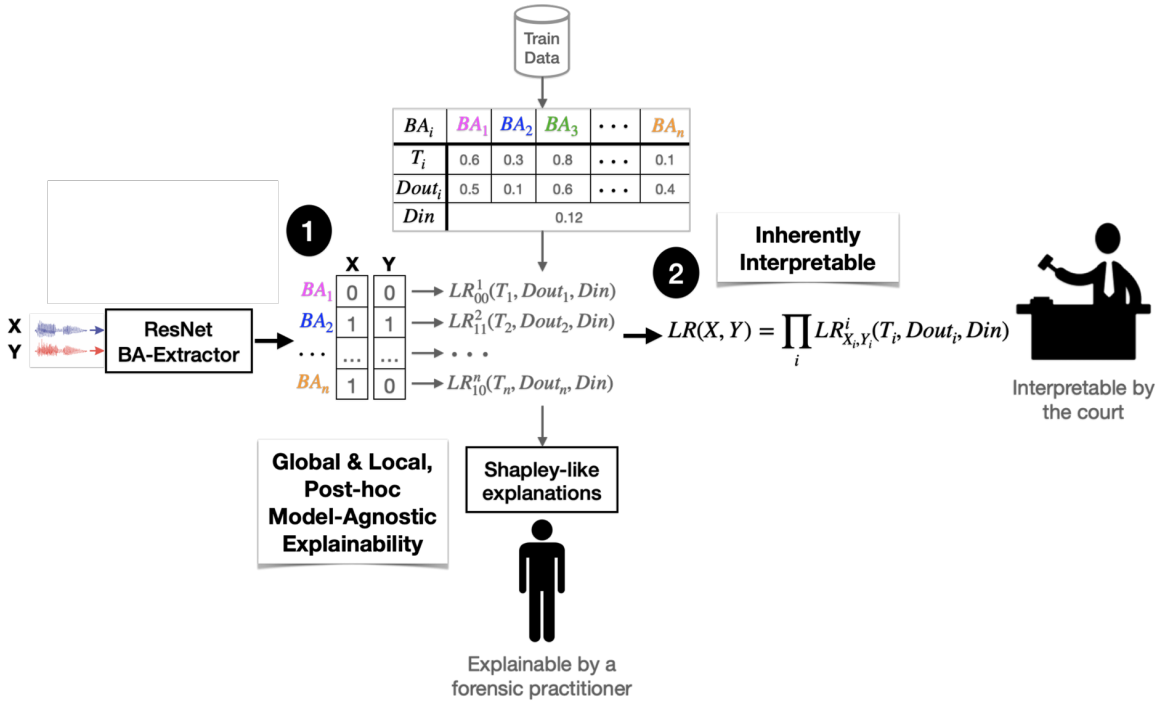


Figure 7.12: Overview of Step 1 & 2 of our approach with interpretability and explainability aspects

Figure 7.12 summarizes an overview of the two steps, along with the added explainability and interpretability aspects. Within the proposed BA-LR framework, the initial step involves extracting binary-attribute-based vectors from two given speech samples. Subsequently, the LR is computed as a product of factors, each dedicated to a specific speech attribute. The behavior of each attribute, characterized by its discriminant power and reliability, is directly involved in the computation of attribute factors, denoted as attribute LRs. These attribute LRs are calculated, with two versions, based on prosecution and defense hypotheses derived from a forensic context, demonstrating their inherent interpretability and informativeness. In the specific case of forensic science, presenting a LR computation that is fully transparent and interpretable to the court allows a better handling of the value of evidence and facilitates the decision-making by judges[5]. Supplementing this LR with details about the most influential attributes, their behaviors, and their impact on the final value ensures a reliable and explainable quantification of the evidence. Even though the BA-LR approach is initially designed based on forensic principles for LR calculation, its applicability is not restricted to forensic science. Rather, it is crucial to mention that it extends

[5]Which is one aspect, the other aspect is relying also on the quality assurance, validation and accreditation of the method

beyond forensic science to encompass general speaker recognition tasks.

Below, we delve into a more detailed discussion regarding the advantages and limitations concerning each phase of BA-LR approach. Additionally, we outline some improvements that may be addressed later or considered in future perspectives.

Advantages of BA-LR approach

Our BA-LR approach offers several significant advantages, which can be summarized as follows:

- *Easy to understand behavioral parameters*: The attributes are characterised in our approach by behavioral parameters that are easy to understand for a human. We proposed in this chapter a basic and a straightforward estimation of these parameters. This initial estimation draws inspiration from the DNA process, intentionally prioritizing simplicity over robustness to offer an easily understandable calculation, following the definitions.

- *Inherently interpretable LR estimation*: The attribute LRs composing the final LR are **inherently interpretable**. Their calculation takes into account both the behavioral parameters of the attribute and the presence or absence of the attribute. This quantification is more reliable than a mere cosine similarity, which compacts all information from both necessary and unnecessary dimensions into a single abstract score. In a forensic context, these interpretable attribute LRs are inherently persuasive and highly appreciated by the court, since they are computed based on prosecution and defense hypotheses.

- *Proposal of two BA-LR versions*: The attribute LRs are computed using two versions. A first naive version that was inspired directly from DNA individualisation. A second version, more logically adapted to the case of voice comparison, provides automatically similar formulations for the case 01 or 10. This second formulation is also shown to correct the distribution of $LLR_{01|10}$ pushing them to be almost negative and therefore more convincing.

- *Discrimination and generalisation abilities*: BA-LR estimation is evaluated for a speaker verification task on three different test datasets having no overlap in terms of speakers with the train dataset. It is essential to recall that the behavioral parameters used to calculate the attribute LRs are estimated on the train dataset and used for the evaluation. Based on that, our solution generally demonstrates discrimination abilities, using both versions, on all three datasets. Compared to the baseline, it shows a slight absolute loss of 1.72% and 1.96%, for DNA-inspired and Speech-adapted, respectively. Given that it uses a $\sim$40 times more compressed speaker embeddings, this loss is deemed acceptable with respect to the well-known trade-off between performance and explainability [181, 247]. These results lead us to firmly believe that our proposal possesses a high potential to build a fully interpretable speaker recognition framework. This potential is further validated and reinforced through the results obtained with a new version of the BA-extractor provided in Chapter 10.

- *Inherently explainable LR estimation*: One key advantage of BA-LR framework, is that attribute LRs share similar properties with Shapley values, allowing their use accordingly (Figure 7.12). This makes the BA-LR approach **inherently explainable** as well, reducing the need for additional methods and saving both time and effort for the forensic practitioner. These explanations are provided at the local level, for each single comparison pair. They are as well provided at the global level by averaging the contribution of attributes across all comparison pairs, while reporting the overall aspect of LR calculation.

Self-criticism and perspectives of BA-LR approach

BA-LR approach presents an interpretable and explainable system for speaker recognition task. So far, it is shown to provide satisfactory results. Nevertheless, as an initial attempt to implement this approach, we are aware that the solution presented in this chapter may exhibit some shortcomings and requires additional improvements. In the following, we discuss these limits from a critical perspective, and suggest avenues for improvement:

- *The notion of speaker profile is misleading*: As previously mentioned in §7.3, the concept of a speaker profile inspired by DNA cannot be directly applied to speech data. It is crucial to emphasize that obtaining ground truth information about the speaker profile, where a speaker profile contains all voice characteristics of a speaker, is inherently challenging. Indeed, speaker profile notion represents the most delicate and challenging foundation of our approach.

- *Limits of the proposed speaker profile estimation*: The method used to calculate the speaker profile in this solution was notably optimistic, impacting the subsequent estimation of behavioral parameters built upon it. A speaker with 10 utterances of variable lengths should not be considered similarly to a speaker having 100 utterances. Additionally, linguistic content represents another significant variability affecting the speaker profile. For example, a speaker distinguished by the distinct pronunciation of an infrequent phoneme may possess only a single utterance featuring this phoneme. As a result, this would increase the drop-out probability of this attribute. Speaker profile estimation is certainly not straightforward, and further studies are required to delve into this direction. As perspective, the speaker profile could be better modeled using a fixed number of utterances having all the same duration and chosen based on phonemic content.

- *Limits of the proposed estimation of behavioral parameters*: Indeed, the estimation of typicality and dropout is easy to understand, but we believe that it is not as robust as expected. The fact that these parameters are based on the speaker profile would in some way or another make their values higher. We acknowledge that averaging the drop-out values calculated per speaker to derive the final drop-out per attribute overlooks the variations in the number of utterances and the variable length of the utterances among speakers. We suggest that the estimation of these behavioral parameters could be improved by employing more sophisticated statistical processes that consider both the number and length of

speech excerpts per speaker. Another suggestion could be to model attributes using Bernoulli-beta distribution using, for a given attribute, the number of ones per each speaker.

- *Final LLR estimation*: The calculation of the LLR as the simple sum of attribute LLRs might be a source of information loss. Even though we experimentally demonstrated in the previous chapter (see §6.4.3) that the dependence between attributes is very low, we agree that it is not perfect. One potential solution to address this imperfection would be to involve assigning weights to the attribute LLRs during the computation of the final LR. This might contribute to balancing the existing dependence and, concurrently, improve the overall calibration of LRs. This solution will be specifically applied in chapter 9.

- *Non calibrated LLRs*: The results provided with the two versions of BA-LR in terms of $C_{llrmin/act}$ demonstrate that the obtained LLR scores are miscalibrated. This is crucial when using LLRs, since non calibrated LLRs are not interpretable as real LLRs. Thus, a further step of calibration is highly required in this case. This is investigated in more details in chapter 9.

- *Trade-off performance Vs. interpretability*: While our BA-LR approach is applicable to any general speaker recognition task, and not limited solely to forensic contexts, the established trade-off between performance and interpretability might not seem convincing for ASpR systems unless coupled with applications in critical fields, like forensic science.

Indeed, so far, the level of explainability and interpretability of the whole approach is not yet fully achieved. An unanswered and legitimate question that may arise at the end of this chapter is: What precisely do these attributes represent in terms of voice characteristics? Or, what is the nature of these attributes and the information encoded within them? This will be effectively addressed in the next step of our approach, namely Step 3.

Chapter 8

Step 3: Attribute explainability

8.1 Introduction . . . 110
8.2 The three-world explainability method . . . 111
8.3 Utterance-level: attribute phonetic description . . . 112
8.3.1 Methodology . . . 112
8.3.2 Inherently interpretable classifier . . . 114
8.3.3 Statistical test . . . 115
8.4 Experiments and results . . . 116
8.4.1 Setup . . . 116
8.4.2 Discrimination ability of the attribute model . . . 118
8.4.3 Attribute explainability in terms of phonetics . . . 119
8.5 Frame-level: attribute phonemic and temporal description . . . 123
8.5.1 Attribute related frame-level information . . . 123
8.5.2 Attribute explainability in terms of phonemes . . . 126
8.5.3 Attribute explainability in terms of localized temporal information . . . 129
8.6 Discussion and perspectives . . . 132

So far, we have characterized the attributes within the BA-LR framework based on their discriminant power and reliability. However, we currently lack information regarding the vocal characteristics of these attributes. In this chapter, our objective is to provide a description of the attributes, shedding light on their nature in terms of acoustic, phonetic and phonemic aspects. This third step would enhance the explainability level of our approach, offering better insights about which information is encoded in the BA-vectors.

8.1 Introduction

The BA-LR scoring framework has demonstrated its potential to effectively enhance interpretability in ASpR systems. However, the employed BA-extractor, discussed in Step 1, adopts a bottom-up approach to extract the binary-attribute-based embeddings. This extractor provides no information or labels associated to the attributes within the embeddings. In the literature, explaining the information encoded within speech representations poses a considerable challenge, especially when these embeddings originate from complex architectures.

Most of the existing research commonly employs probing classifiers to illustrate the presence of specific predefined speaker characteristics within the representations. For instance, the work in [248] delved into the encoding of various properties, including speaker identity, gender, speaking rate, text content, and channel information, within speaker embeddings. The authors in [249] adopt the same approach to reveal information related to the speaker, channel, transcription (i.e. sentence,words, phones), and meta information about the utterance (i.e. duration and augmentation type) from speaker embeddings. In the context of an automatic speech recognition task, the work in [250] demonstrates how information such as speech style, accent and broadcast type are encoded through the layers of the neural networks. Concurrently, the work in [251] shows how accent information is reflected in the internal representation of speech. In a dialect identification task, the work in [252, 253] explores the encoding of non-dialectal information within the model, comparing it with dialectal information through probing classifiers for gender, voice identity, languages, and channel quality. Notably, all these works are based on finely labeled data to perform the supervised probing classification, which is a resource that is both critical and scarcely available due to its associated costs.

In a different avenue, some other works analyzed the presence of phonemic information along neural network layers [254, 255, 256]. For a speaker recognition task, the work in [257] extracts a frame-level representation from each layer of the neural network and studied the encoding of phonemes as well as the phonetic classes to investigate the functioning of the speaker embedding model. A recent study by [258] focuses on the presence of acoustic features (i.e. F0, intensity, duration, formants...etc) in model layers. It explored how the network encodes this information by establishing relationships between these variables and activations across the intermediate convolutional layers of the model.

In this chapter, our objective is to explain the binary-attribute-based embeddings. More specifically, we aim to describe the nature of the attributes extracted in our approach without relying on any form of annotation. To fulfill this objective, we introduce a novel explainability methodology that enables automatic description of attributes in terms of any type of available information. This represents the third step of our three-step approach. In the next sections, we start with an overview of this methodology. Then, we propose two levels of attribute description: an utterance-level and a frame-level. For the former, we detail the applied methodology then we present our findings quantifying the effectiveness and the fidelity of this phonetic description in accurately

representing the nature of the attributes. The explanations are provided on both an individual attribute basis and collectively for groups of attributes that share common phonetic information. For the latter, we describe the association of each attribute to frame-level information, then we use this information to provide an explanation using phonemes, classes of phonemes and the localized temporal information. In summary, we conclude this chapter with a discussion of our findings.

8.2 The three-world explainability method

In this section, we introduce the core concept of the three-world explainability method, building upon the description of attributes. Figure 8.1 illustrates a description of this method, as an interaction between three worlds defined as follows:

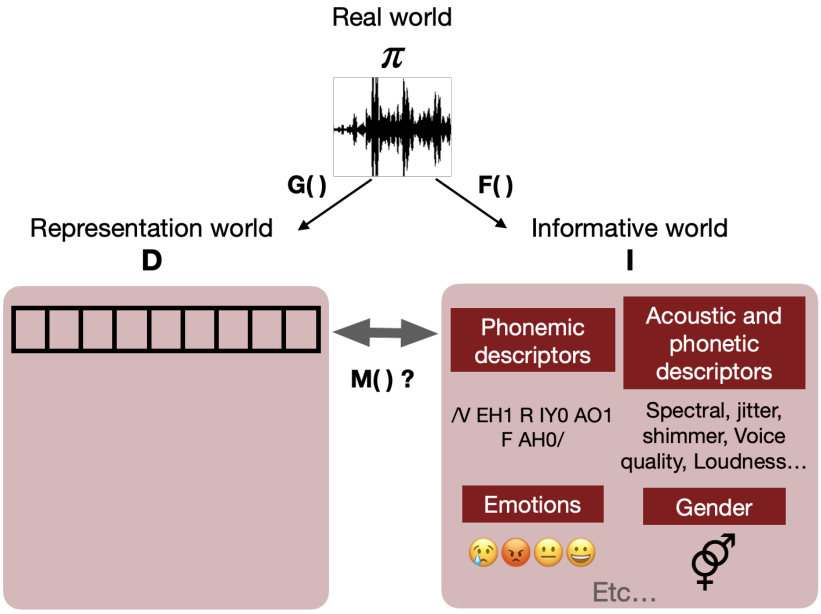


Figure 8.1: The three-world illustration of our proposed methodology

- A *Real world (π)* representing a speech extract, from which it is possible to derive two different worlds using two functions $G()$ and $F()$.
- A *Representation world (D)* which illustrates a high-dimensional discrete or continuous representation of the real world (e.g. BA-vectors). This representation is typically extracted using a DNN model denoted by the function $G()$.
- An *Informative world (I)* that contains all available information about the real world (e.g. Emotions, gender, phonemes...etc) as well as information that is directly extracted from the real world (e.g. pitch, formants, jitter,...etc). The extraction of this information is generally and ideally performed by a human annotator. But it could be also done using automatic tools denoted by the function $F()$.

Following this idea, a central question arises: **How to determine a function $M()$ that establishes a mapping between D and I?**

Indeed, the determination of this mapping would serve as a way to explain the encoded information in D in terms of the available information in I. This mapping could be any function that establishes a relationship between I information and D representations. It could be seen as a classification problem where I are features used to predict specific dimensions of D (i.e. 0 or 1) seen as target, or it could be some statistical or information-theory techniques that establishes this relationship. The key advantage of this method is its flexibility. Even in the absence of labels and annotations, explainability of the D world could be performed from the π data itself.

8.3 Utterance-level: attribute phonetic description

The extracted attributes of the speaker embeddings currently lack information related to voice characteristics. In this section, we build upon the three-world explainability method to introduce a novel methodology specifically designed to automatically describe the inherent nature of attributes. We begin by outlining the prerequisites to be met. Following that, we introduce the main idea and foundations of the proposed approach. Subsequently, we delve into the detailed description of each sub-step of this methodology.

Prerequisites

The prerequisites we set for this methodology are the following:

- To explain attributes deriving from a bottom-up BA-extractor.
- To not require labels or annotations or any additional manual labeling.
- To produce an automatic and accurate description.

8.3.1 Methodology

Based on the-three world methodology, we formulate the following definitions, referring to our requirements:

- The π world is defined by speech extracts.
- The D world is defined by the binary-attribute-based embeddings, namely the BA-vectors.
- The $G()$ function represents our BA-extractor that extracts the BA-vectors.
- The I world is represented by acoustic, phonetic, phonemic and temporal information, namely *Descriptive variables*.

- The $F()$ function represents any automatic tool that is able to automatically extract this information directly from the π world.
- The $M()$ function is the mapping to be determined between each attribute in the D world and available information in the I world.

With these definitions in mind and given the binary nature of our D world, our methodology is built under the assumption that:

> ***If we can identify variables in the I world that effectively differentiate between the presence and the absence of a particular attribute in the D world, THEN these variables are likely to be good descriptors of the attribute.***

The proposed methodology follows a three sub-steps strategy applied independently for each attribute BA_i [259], as illustrated in Figure 8.2.

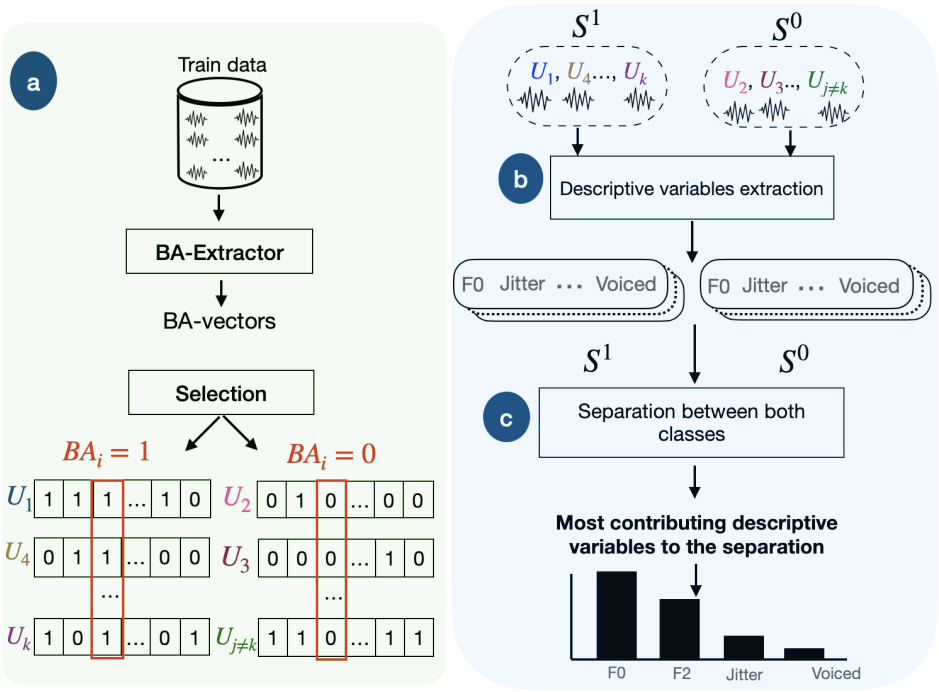


Figure 8.2: Methodology of an attribute explainability following (a), (b), and (c) sub-steps, as applied to each attribute

a) Selection of speech extracts

Using the D world, the speech extracts of the train dataset in the π world, are grouped into two sets as depicted in Figure 8.2-a. The first set, denoted "S_1", groups the extracts where the considered attribute, BA_i, is present. The second set, denoted "S_0", contains the extracts from **speakers other than those present in S_1**, and obviously

where the attribute has a value of 0. In other words, S_1 contains positive examples of the attribute pronounced by the set of speakers who share this attribute, while S_0 presents negative examples pronounced by other speakers, who never had the attribute. The intentional absence of speaker overlap between the two sets is designed to not be influenced by drop-out phenomenon (refer to §7.3.2). Finally, some randomly selected extracts are eliminated from S_0 to balance the number of extracts in the two sets in order to avoid bias during the selection process.

b) Extraction of descriptive variables

The second sub-step, in Figure 8.2-b, is dedicated to extract information of I world from the π world. For this purpose, we choose a set of descriptive variables of the I world, which can be of any type, as long as they can be computed automatically from speech. For this specific work, we opt to use descriptive phonetic variables[1], but any other variable type ,available in the I world, is possible, such as available annotation, phonemic [260, 261, 262] or language-related variables [263]. The values of the variables are then extracted for each speech extract.

c) Mapping function

The last sub-step, in Figure. 8.2-c, is to determine the mapping function $M()$ between the descriptive variables of the I world and the attribute values in the D world. The strategy relies on determining the variables that best explain the difference between sets S_1 and S_0 of the attribute. For this purpose, we propose two solutions. The first one consists in training an inherently interpretable classifier to separate examples of sets S_1 and S_0. It uses the descriptive variables extracted from each training example as features. The most influential variables are the ones that best describe the attribute in question. To ensure the relevance of this choice and to have a comparison basis, we use a second solution based on a statistical test, which is less powerful, but provides a simpler solution. Both solutions are further detailed in the remaining of this section.

8.3.2 Inherently interpretable classifier

The goal of this first method is to find a model that is interpretable by nature and able to discriminate between the two classes of the attribute using the descriptive variables as features. The idea of using an inherently interpretable model to predict the black box model output and explains its predictions, is inspired from surrogate models (refer to §4.4.3). The ideal candidates are those inherently interpretable models, capable of giving sufficient separability performance between class 1 (i.e. S_1) and class 0 (i.e. S_0) examples. Additionally, they should be easy to train, fast and characterized by stability and minimal computational costs. In line with these criteria, we opt for a Decisiontree, given its reputation for speed, simplicity, and inherent interpretability.

[1] such as F0, formants, jitter, shimmmer, etc.

This Decisiontree model serves to verify whether the descriptive variables, taken as features, are able to separate the two classes of the attribute as illustrated in Figure 8.3.

Following this, if the model proves an acceptable separability performance between the two classes for the attribute, then a further step is needed to retrieve the most contributing descriptive variables to this separation. To this end, we use TreeExplainer from SHAP toolkit[2], which computes Shapley values [199] adapted for Decisiontree models. This variant of SHAP is fast and efficient especially when applied on Decisiontree models.

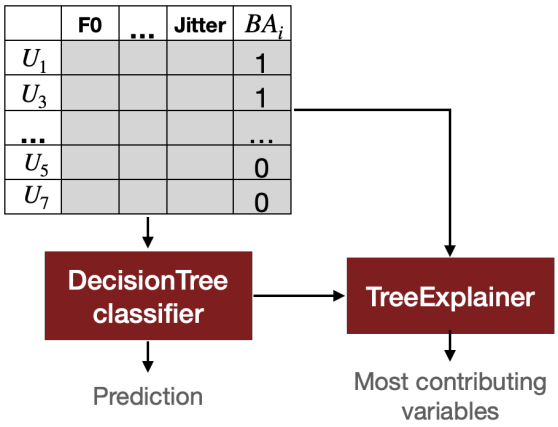


Figure 8.3: Application of the Decisiontree classifier for Attribute BA_i with TreeExplainer from SHAP

Shapley values are calculated following Equation (4.1) and they are used to estimate the average feature contribution to the classifier predictions following Equation (8.1), where X_j is a given descriptive variable, BA_i is the attribute described, M is the number of descriptive variables and $ShapMean(X_j)$ is the average of Shapley values obtained for X_j across all instances using the BA_i model.

$$\text{Contribution}_{BA_i}(X_j) = \frac{ShapMean(X_j)}{\sum_{k=1}^{M}(ShapMean(X_k))} \tag{8.1}$$

8.3.3 Statistical test

The second method consists in a statistical test that selects a subset of the most important variables to separate the two classes. In this work, we propose to use Stepwise Linear Discriminant Analysis (SLDA) [264, 265] as a statistical method that identifies a linear combination of the descriptive variables that most discriminate the examples of the two classes 1 and 0. Feature selection is based on the *Wilk's Lambda* criterion[3] as expressed in Equation (8.2), where *det* is the determinant, A is the within class

[2]https://github.com/slundberg/shap

[3]https://www.blackwellpublishing.com/specialarticles/jcn_9_381.pdf

covariance matrix, and B is the between class covariance matrix.

$$\text{Wilk's lambda} = \frac{det(A)}{det(A+B)} \tag{8.2}$$

Wilk's Lambda is calculated for each descriptive variable and reflects the discriminant power of the variable. The selection procedure follows a systematic step-by-step approach. Initially, the algorithm identifies the variable with the highest discriminant power for the classes until Wilk's Lambda achieves statistical significance. In subsequent steps, the model is reassessed: the variable not currently in the model but contributing the most to the discriminant power is incorporated. Among all variables already in the model, if the variable contributing the least to the discriminant power, as measured by Wilk's lambda, exceeds the significance threshold, set to 0.01 [**<empty citation>**], the variable is removed. This iterative process continues until all variables are tested.

8.4 Experiments and results

In this section, we conduct experiments to validate the proposed methodology. We begin by setting up the datasets, followed by detailing the extraction of descriptive variables and configuring the models. Next, we demonstrate the discrimination capabilities of the inherently interpretable models associated to the attributes. This discrimination ability enables us to subsequently identify the most influential variables contributing to this differentiation. Following this, we provide an automatic description of an example of an attribute, explaining its nature in terms of acoustic and phonetic information. This description is strengthened and validated through the findings of the statistical method. Expanding our perspective, we demonstrate that certain groups of attributes share common phonetic and acoustic aspects. Finally, this description is evaluated to show its robustness and accuracy.

8.4.1 Setup

Datasets

In this experiment, we use two datasets:

- The train dataset is the same as the training dataset of the BA-extractor, namely VoxCeleb2 (Vox2) (§6.4.1). This choice is driven by the assumption that attribute information is more robust and present under the training dataset. This is because the extractor was primarily trained on this dataset for the extraction of these attributes. This dataset is divided into two subsets S_0 and S_1 for each attribute, then used to train the mapping functions. This yields 205 S_0 and S_1 sets of utterances from train data.

- A test dataset, namely VoxCeleb1 (Vox1) (§6.4.1), is employed in this experiment to evaluate the stability and the fidelity of the information learned by the inherently interpretable classifiers. This dataset is also divided into two sets S_0 and S_1 for each attribute. We recall that there is no intersection in terms of speakers between both datasets.

Following the first sub-step, BA-vectors are extracted and speech samples are selected to form two sets S_0 and S_1 per attribute for both datasets. Further details regarding the number of speech extracts chosen in both train and test datasets for each attribute are provided in Appendix C in Table C.1.

Descriptive variable extraction

In order to extract the descriptive variables, we opt to use OpenSmile[4] toolkit, an open-source audio feature extraction toolkit. We specifically use the pre-defined set of variables *eGeMAPS*[5] [266], which contains 25 low-level descriptors (LLD), extracted at the frame level, grouped as follows:

- **Frequency related parameters** such as the pitch (F0), Jitter and Formant 1, 2, and 3 (F1, F2, F3) frequency and bandwidth.
- **Energy or Amplitude related parameters** such as the shimmer, Loudness, Harmonics-to-Noise Ratio (HNR).
- **Spectral parameters** such as Alpha Ratio, Hammarberg Index, Spectral Slope, Formant 1, 2, and 3 relative energy, Harmonic difference H1-H2, Harmonic difference H1–A3, MFCC 1–4 Mel-Frequency Cepstral Coefficients and Spectral flux difference of the spectra of two consecutive frames.

All LLD are averaged over time. Arithmetic mean and standard deviation are applied as functionals to all LLD. 8 additional functionals are applied to loudness and pitch: 20-th, 50-th, and 80-th percentile, the range of 20-th to 80-th percentile, and the mean and standard deviation of the slope of rising/falling signal parts. Added to that, 6 temporal features are incorporated such as the rate of loudness peaks, the mean length and the standard deviation of both continuously voiced regions and unvoiced regions and the number of continuous voiced regions per second. This finally yields 88 parameters. Further details regarding the definitions of these parameters can be found in [266]. The extraction of descriptive variables is performed for both the train and test datasets.

Decisiontree and SLDA configuration

After selecting train sets S_0 and S_1 of 88 descriptive variables for a specific attribute, we begin by standardizing the training data by centering on the mean and scaling to

[4]https://github.com/audeering/opensmile

[5]In this work, the set of parameters is arbitrary chosen.

unit variance. Subsequently, we train the Decisiontree classifier, with a key parameter being Max_depth, indicating the depth of the tree. To determine the optimal tree depth for each attribute, we conduct a grid search to identify the value that enhances training accuracy. Further details about the number of samples in train and test for each model as well as models configuration are provided in Table C.1. Simultaneously, we utilize the same standardized data for the statistical test. As previously explained in §8.3.3, the SLDA algorithm requires only a significance threshold to stop the addition of variables. This threshold is consistently set at 0.01 for all attributes.

8.4.2 Discrimination ability of the attribute model

In this section, we use the inherently interpretable model solution in order to build automatic descriptions of the different attributes in terms of descriptive variables. The three sub-steps process previously described (§8.3.1) is performed independently for each attribute. An attribute model is trained to classify between 0 and 1 classes using the 88 descriptive variables of the train data. Then, the model is evaluated on speech samples of the test data selected with respect to the concerned attribute using the same process. For further details, refer to AppendixC in Table C.1.

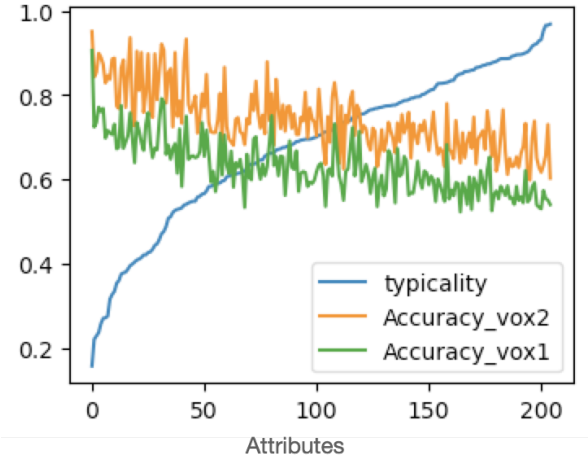


Figure 8.4: Accuracy of attribute models on Vox2 (train) and Vox1 (test) along with their associated typicality values.

Figure 8.4 shows the accuracy values obtained for all attribute models on both the train (i.e. VoxCeleb2) and test datasets (i.e. VoxCeleb1), ranked by the typicality of the BAs, from lowest to highest. The attribute models exhibit relatively high accuracy values on the training set, ranging from 0.6 to 0.97. This demonstrates the discrimination capability of each attribute model. The fact that these models are trained using the set of descriptive variables signifies that these variables are effective in distinguishing between the two classes of the attribute. As can be noticed, some attribute models are shown to be better than others. This could be explained by the fact that the descriptive variables are effectively held by these attributes. The slight difference in

accuracy values, averaging 0.11, between the train and test datasets not only suggests well-chosen and well-fitted models but also highlights the stability and fidelity of the information conveyed by the attributes in relation to descriptive variables.

The rationale behind adding the typicality of the attribute is mainly to establish a relationship between the behavior of the attribute and the robustness of information held by the attribute. Figure 8.4 shows a strong inverse correlation between attribute typicality and accuracy, indicating that the attributes that carry the greatest power to discriminate between speakers are also the best represented by the descriptive variables.

8.4.3 Attribute explainability in terms of phonetics

In this section, we present the explainability of an attribute following three phases. 1) We describe an attribute by the most important descriptors, selected by the inherently interpretable model and SLDA, demonstrating a convergence between results of both methods. 2) We evaluate this attribute description by training 88 inherently interpretable sub-models, using each time, increasing number of the most important descriptive variables. 3) We provide a global description grouping all attributes and showing the corresponding contribution of the families of variables.

Here, we choose to investigate 1 and 2 phases of this process for one example attribute. We opt for BA_9 attribute because of its high discrimination capability. The results for the other attributes can be found in our GitHub repository[6].

Example of attribute BA_9 description

Given the discrimination capability of the attribute classifier, TreeExplainer is applied to select the most contributing descriptive variables of the attributes. In Figure 8.7, the contributions of descriptive variables to BA_9 are illustrated, organized by families. These families represent the LLD descriptors composed of groups of functionals as depicted in the figure. As observed, a functional of the F_0 family stands out as the primary contributor, accompanied by other less contributing functionals associated with F1 and F2 formants.

To establish a comparison and reinforce this description, we apply the SLDA method for the same attribute example, BA_9. It is important to recall that SLDA method selects descriptive variables based on the Wilk's Lambda which quantifies the discriminant power of the variable. Figure 8.5 shows the Lambda values as a function of the selected number of variables. A closer look into the 10 most discriminant variables reveals certain similarities with the description offered by the attribute model in Figure 8.7.

[6]https://github.com/LIAvignon/BA-LR/tree/main/Step3/explainability_results/Explainability

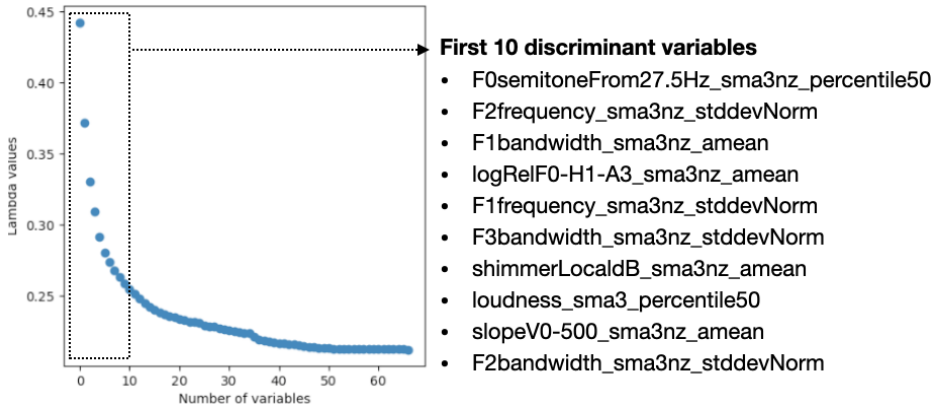


Figure 8.5: Wilk's Lambda values as a function of the 65 selected variables for attribute BA_9, with a closer look on the first 10 variables

To measure the similarity between variables identified by both methods, we choose the variables from the attribute classifier that collectively represent 75% of the contribution and we assess their intersection with the variables selected by the SLDA method. We obtain approximately 80% convergence between both methods. This enhances our confidence in the obtained attribute description.

Evaluation of attribute BA_9 description

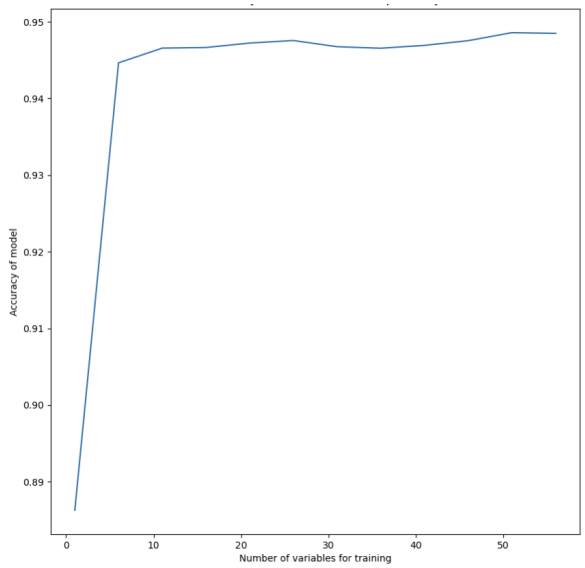


Figure 8.6: Evolution of accuracy of sub-models of BA_9, each trained with incremental number of descriptive variables: from most to least contributive

Figure 8.7: Descriptive variables contributions to the BA_9 model, grouped by families

Thus far, we have employed two distinct methods to describe the attribute, and remarkably, we have obtained convergent results between the two approaches. This convergence not only serves as a means of verification but also strengthens the resulting attribute description. To further evaluate this attribute description, we conduct a dedicated experiment on this attribute, namely BA_9.

In this experiment, we train 88 sub-models on the training dataset, each configured identically to the attribute model, using an incremental number of descriptive variables—from the most to the least contributing. To elaborate, the first sub-model is trained with only the most contributing descriptive variable, the second with the two most contributing, the third with the three most contributing, and so forth. Figure 8.6 shows the evolution of the accuracy of these sub-models trained with incremental number of descriptors. Remarkably, even with only the most contributing (Figure 8.7) and discriminant variable (Figure 8.5), i.e "F0semitoneFrom27.5Hz_sma3nz_percentile50", the accuracy of the first sub-model in classifying between the 0 and 1 classes of BA_9 is approximately 88.7%. With only the five most contributing variables, the fifth sub-model achieves an accuracy of 94.5%. This demonstrates the reliability and accuracy of the provided attribute description.

Description per group of attributes

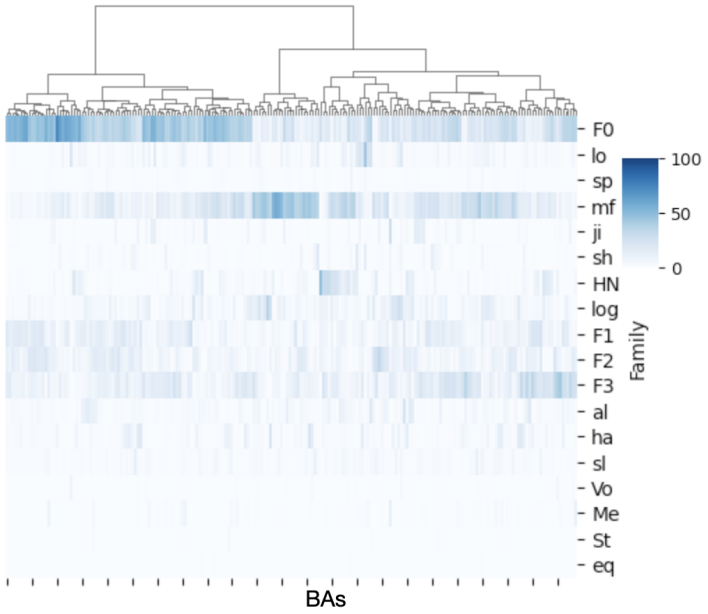


Figure 8.8: Heatmap illustrating the contribution of variable families to attributes grouped into clusters

From a broader perspective, we calculate the average contributions of descriptive variables per family across all attributes using Equation (8.1). Figure 8.8 illustrates a shared phonetic description among groups of attributes in terms of families of variables. The definition of these families is provided in Figure 8.7. The attributes in

Figure 8.8 are ordered by hierarchical clustering with single linkage. The differences between the attributes (i.e, BAs along the x-axis) reinforce the hypothesis that the attributes encode different phonetic information. Notably, certain variable families, such as the F0 family and the mf family (i.e., MFCCs), exhibit a more pronounced emphasis on average. This is evident in the concentration of contribution within these families for certain groups of attributes, as opposed to others where the overall contribution is more evenly distributed among all families.

8.5 Frame-level: attribute phonemic and temporal description

This section is dedicated to provide a description of attributes in terms of phonemes and a localization of temporal information. Our objective is to introduce an additional dimension to the previous description, providing a more comprehensive understanding of its nature. For this end, we employ a post-hoc, model-specific explainability methodology. The concept involves retracing the flow of information associated with a specific attribute through the layers of the DNN architecture of the BA-extractor until reaching the input. This process enables us to associate each attribute with its frame-level information and to explore a finer-grained level of description, following the three-world methodology.

Research questions

In this section, our focus is on addressing the following questions:

- How can we localize and extract frame-level information associated with a specific attribute?
- Can we establish an alignment between attributes and input frames?
- Are there attributes that specifically encode certain phonemes? Is there a distinct class of phonemes consistently encoded across attributes?
- Do attributes differ in terms of the localized temporal information?

8.5.1 Attribute related frame-level information

In this section, we aim to obtain an alignment between attributes and frame-level units. To accomplish this, we firstly recall the DNN architecture of the BA-extractor of Step 1 and apply a modification to create a direct access to frame-level information from attributes. Secondly, we localize this information. Finally, we detail the process of retracing back the flow into the DNN model, yielding into an alignment between attributes and input frames.

Revisiting the DNN architecture and modification

It is important to recall that the BA-extractor, discussed in Step 1 (§6.3), is initially trained to extract activation vectors, such as Softplus-vectors. These vectors are binarized after training to obtain the BA-vectors. Figure 8.9 revisits the architecture of the BA-extractor, which takes N_f frames of 61 filterbank outputs as input. The architecture comprises a ResNet extractor that produces temporal units of 2048 dimensions flattened over 256 filters, each of 8 dimensions. Progressing to the utterance-level, a pooling layer averages information across all temporal units, succeeded by a fully connected layer (FC) and a Softplus activation, yielding the Softplus-vector.

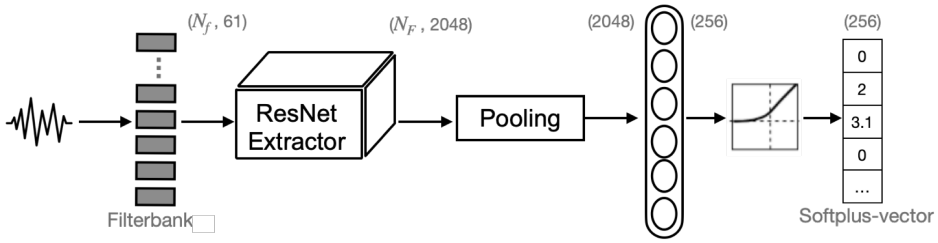


Figure 8.9: DNN architecture of the BA-extractor

Figure 8.10 illustrates the modification on the BA-extractor to represent attributes with frame-level units. The key modification is to eliminate the pooling layer from the network. The output frame-level units from the ResNet extractor are denoted as **MegaFrames (MF)**. The FC layer with Softplus activation is added after the ResNet's output following Equation (8.3). Here, M represents the matrix output from the ResNet extractor, sized N_F multiplied by 2048, where N_F denotes the number of MegaFrames. The matrix W sized 2048*256 and vector b sized 256 comprise weights and biases, respectively, learned during the training of the BA-extractor.

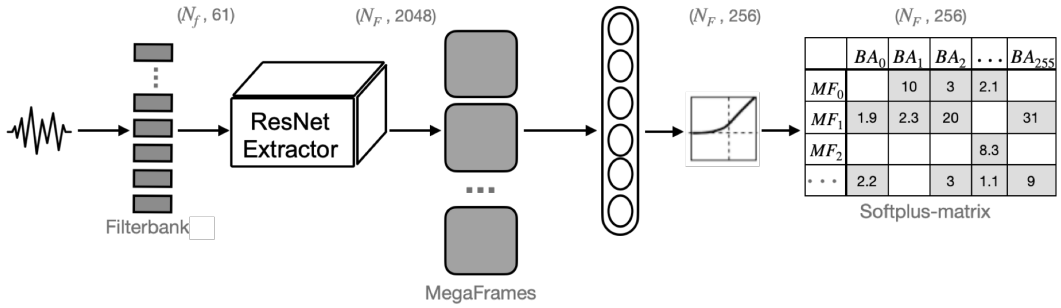


Figure 8.10: Modified DNN architecture for frame-level information extraction

$$\text{Softplus}(\underbrace{M}_{(N_F*2048)} \cdot \underbrace{W^T}_{(2048*256)} + \underbrace{b}_{(256)}) = \underbrace{A}_{(N_F*256)} \tag{8.3}$$

The resulting matrix A, sized N_F*256, corresponds to the *Softplus matrix*, representing MegaFrames activations to BAs, which are 0 or positive value. This matrix enables

an automated selection of active (i.e. non zero) MegaFrames for each BA. So far, for a given input utterance, we can obtain two distinct activation representations: 1) an utterance-level representation, which is the Softplus-vector, and 2) a frame-level representation, which is the Softplus-matrix. Further analyses of these representations are provided in Appendix C.

Localizing frame-level information through the network

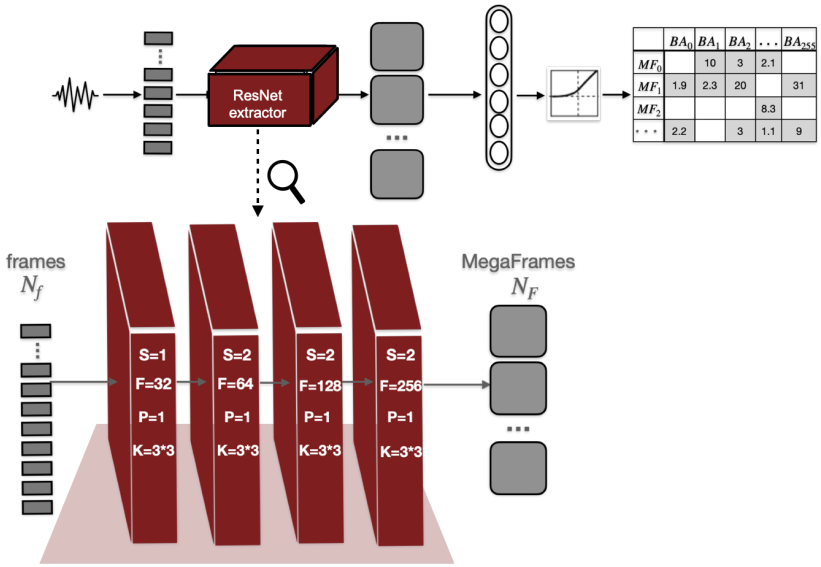


Figure 8.11: A closer look into the ResNet extractor illustrating the relationship between the input frames and the output MegaFrames through ResNet blocks.

Examining the ResNet extractor in detail allows us to trace the flow of information from MegaFrames back to the input frames of filterbanks. We recall that the employed ResNet architecture consists of four convolution blocks with a configuration described in Figure 8.11. Here, S denotes the stride, F is the filter, P represents the padding, and K is the kernel size, here, 3. For a given input utterance of N_f frames and the configuration of each block, Equation (8.4) is iteratively applied in each block to determine the number of units at the output of each block. The output of the final block represents the number of MegaFrames, denoted as N_F and expressed in Equation (8.5).

$$Output(I) = \frac{I + 2 * P - K}{S} + 1 \tag{8.4}$$

$$N_F = Output(Output(Output(Output(N_f)))) \tag{8.5}$$

Alignment between attributes and input frames

With this configuration in mind, we can infer that a MegaFrame is represented by 14 frames, with an overlap of 7 frames with the next MegaFrame. This alignment is visually demonstrated in Figure 8.12, where we showcase an example of attribute

alignment. In this specific example, the attribute BA_1, being active for the corresponding utterance, selects MegaFrames MF_0 and MF_1. Examining these two MegaFrames allows us to align the attribute, through the network, with its corresponding input frames from f_0 to f_{22}.

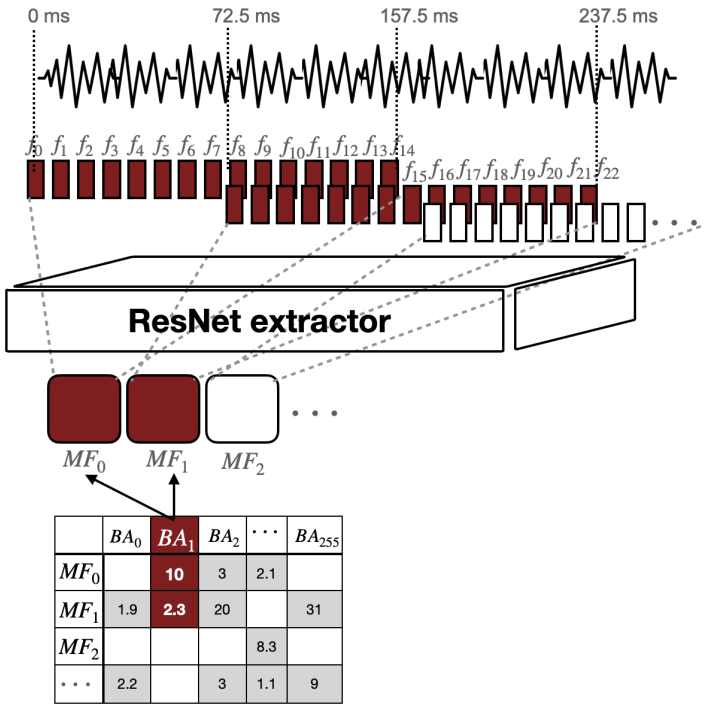


Figure 8.12: An example of attribute alignment with input frames through selected MegaFrames.

8.5.2 Attribute explainability in terms of phonemes

In this section, we aim to describe attributes in terms of phonemes and classes of phonemes. To achieve this, we start by extracting a transcription and an alignment of phonemes for VoxCeleb1 dataset, forming the I world. Then, we show an example of utterance from a single attribute perspective, highlighting the mapping between attribute and selected phonemes. Following that, we employ this mapping to describe each attribute with the number of selected phonemes and classes of phonemes across a set of utterances.

Transcription and phonemes alignment

In the absence of publicly available ground truth transcriptions and phoneme alignments for the VoxCeleb datasets, we use a transcription proposed during JSALT workshop2023[7]. A pre-trained English speech recognition model, in this work WhisperX[8],

[7]Done by the linguistic team from Xdiar project of the JSALT workshop2023
[8]https://github.com/m-bain/whisperX

is used to generate transcriptions with word-level timestamps for a subset of VoxCeleb1 English utterances only. This yields in a subset of only 4822 English utterances transcribed. Subsequently, the Montreal Forced Aligner[9] (MFA) is applied to extract phone boundaries and align them with corresponding timestamps. Table 8.1 presents the employed phonemes grouped by classes following the chart used by MFA.

Table 8.1: Phonemes and corresponding classes based on MFA chart

Class	**Phonemes**
Vowels	AA0, AA1, AA2, AE0, AE1, AE2, AH0, AH1, AH2, AO0, AO1, AO2, AW0, AW1, AW2, AY0, AY1, AY2, EH0, EH1, EH2, ER0, ER1, ER2, EY0, EY1,EY2, IH0, IH1, IH2, IY0, IY1, IY2, OW0, OW1, OW2, OY0, OY1, OY2,UH0, UH1, UH2, UW0, UW1, UW2
Fricative	F, V, TH, DH
Stop	P, B, T, D, K, G
Nasal	M, N, NG
Affricate	CH, JH
Sibilant	S, Z, SH, ZH
Approximant	W, R, Y
Lateral	L

Mapping between attributes and phonemes

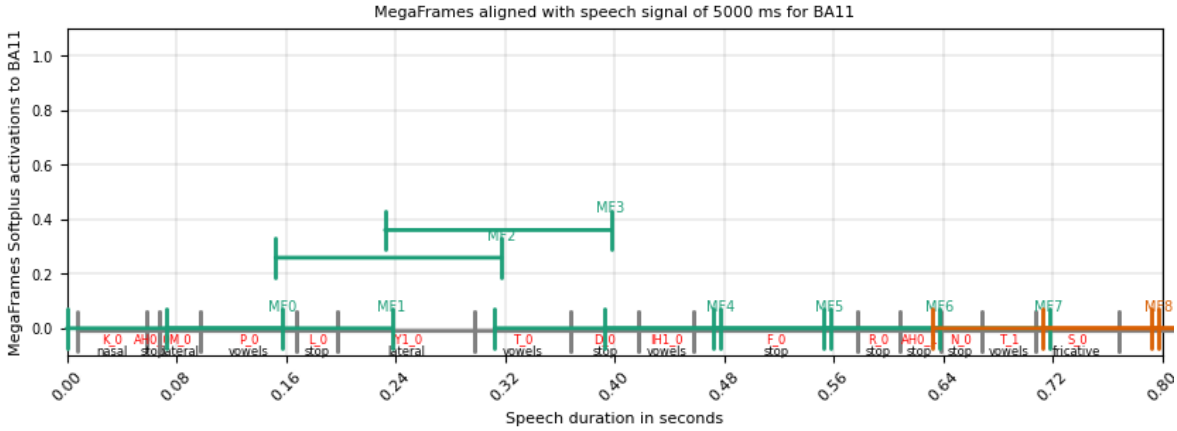


Figure 8.13: Normalized activations of MFs to BA_{11}=1 in a portion of 0.8s of a speech utterance of 5s, aligned with phonemes and classes of phonemes

Figure 8.13 illustrates an example of utterance alignment with phonemes and classes of phonemes along with the normalized activations of MFs corresponding to a specific present attribute (i.e., $BA_{11} = 1$) in the utterance. The normalization of activations is obtained by dividing each MF activation by the maximum MF activation corresponding to a given attribute. As depicted in the figure, the selection of active MFs associated

[9]https://montreal-forced-aligner.readthedocs.io/en/latest/

with the attribute (i.e., MF_2 and MF_3) indirectly establishes a connection between the attribute and the phonemes. Additional examples of alignment for various present attributes in the utterance are provided in Appendix C, §C.3.

Description of attributes in terms of phonemes

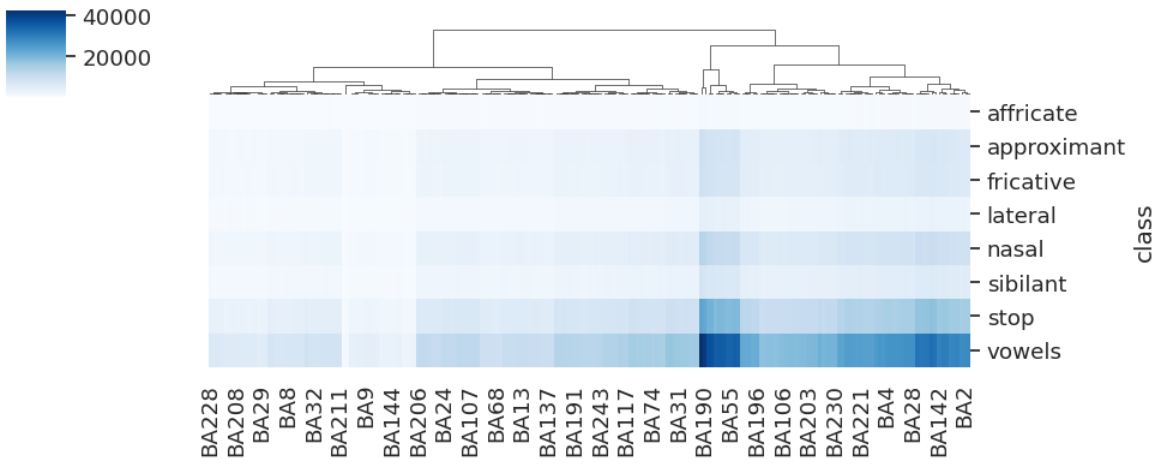


Figure 8.14: Occurrence of each class of phonemes, clustered per BAs

In order to obtain a global description of the attribute in terms of phonemes and classes of phonemes, we need to aggregate phoneme information across all utterances. To achieve this, for each attribute, we count the occurrences of each phoneme and class of phoneme selected by the corresponding active MFs and accumulate these counts across all utterances where the attribute is present (i.e., BA_i=1).

Figure 8.14 demonstrates the occurrences of classes of phonemes per attribute, while grouping attributes sharing similar classes behavior. As can be noticed, the "Vowels" class is the mostly selected by attributes compared to other classes, followed by the "Stops" class and, then "Nasals" class. This result is aligned with other works in [257, 267, 268] for the vowels and nasals, where they are both shown to be important to discriminate speakers. However, for the stop class, this was not expected. Additionally, along the attributes, we can identify some groups that are having a more pronounced selection of classes than others. This is clearly demonstrated for vowels class.

Delving into each class of phonemes, we examine the occurrences of each phoneme to understand whether the class behavior is confined to specific phonemes or is shared among phonemes within the same class. Figure 8.15 illustrates this analysis. In the case of vowels, the phoneme "AH0" stands out as the highly selected by certain BAs compared to all phonemes across all classes (white zones in the heatmap). This selection is similar in terms of BAs but less pronounced for the "IH0" vowel. For stops, the phoneme "T" shows particular importance for some attributes, closely followed by the phoneme "D". In the nasals, it is the phoneme "N" that emerges as highly selected by certain BAs. In the remaining classes, the frequency of phoneme selection by attributes

is less pronounced, yet there are consistently preferred phonemes within each class. For instance, within the fricative class, the "DH" phoneme is more frequently selected than others. Except for the affricate class, where the selection of phonemes is notably low, indicating their relatively lower importance to BAs.

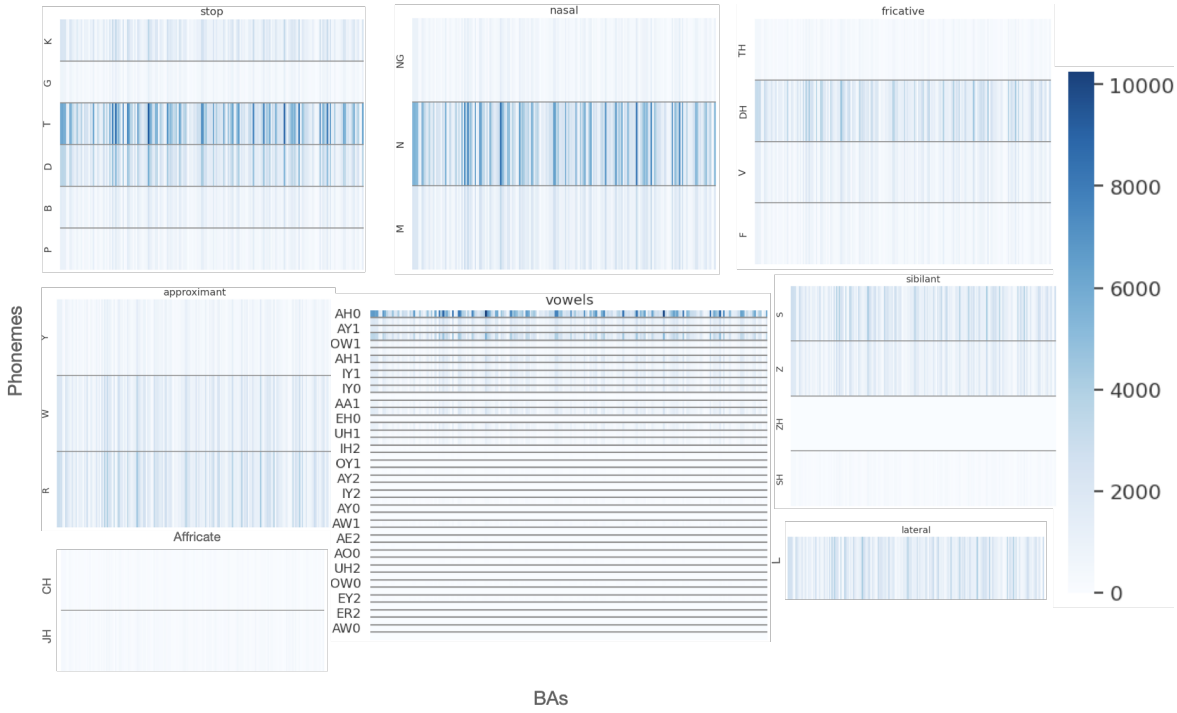


Figure 8.15: Occurrence of each phoneme in its phonetic class

In Figure 8.16, we add a global view of all phonemes together, which clearly shows that there is only some subsets of BAs that particularly selects the vowels, stops and nasals phonemes.

8.5.3 Attribute explainability in terms of localized temporal information

In this section, we are interested in another description of attributes, specifically in terms of localized temporal information. To accomplish this, we rely on the mapping established between attributes and input frames. We briefly describe the process of temporal information localization and extraction per each attribute. Then, we present the description of attributes using this information.

Localized temporal information extraction

The used temporal information in this work, are LLDs extracted at the frame-level for each utterance. These LLD descriptors, described in §8.4.1, are extracted using the same window used for filterbank output extraction in the feature extraction phase of our BA-extractor. Each window is 25ms with 10ms of overlap with the next. This

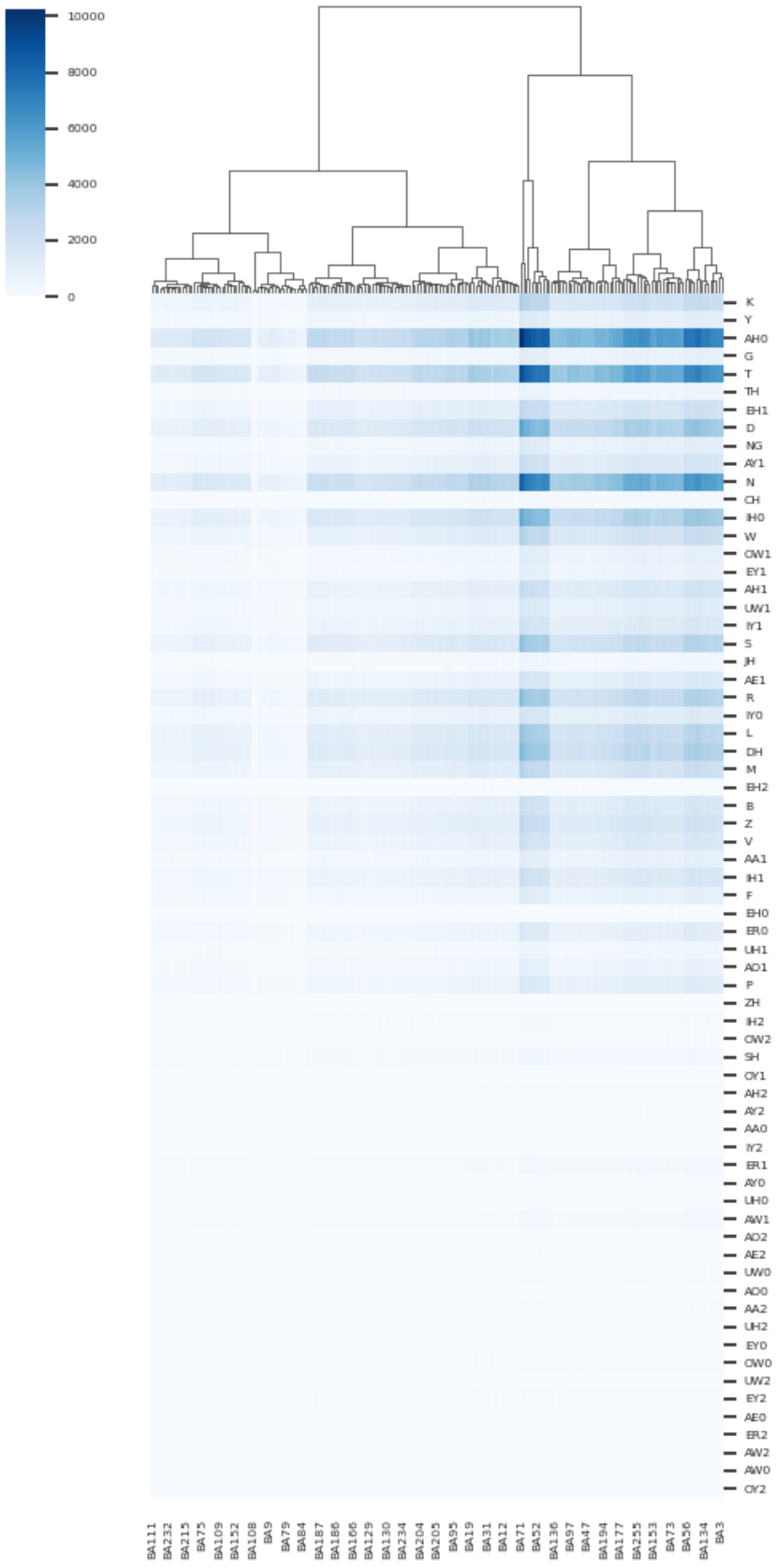


Figure 8.16: Occurrence of all phonemes together, clustered by BA

experiment is also performed on VoxCeleb1 dataset considering for each attribute, the utterances where it is present and the corresponding selected frames. In this setting, we limit the number of frames for each attribute to 2 million.

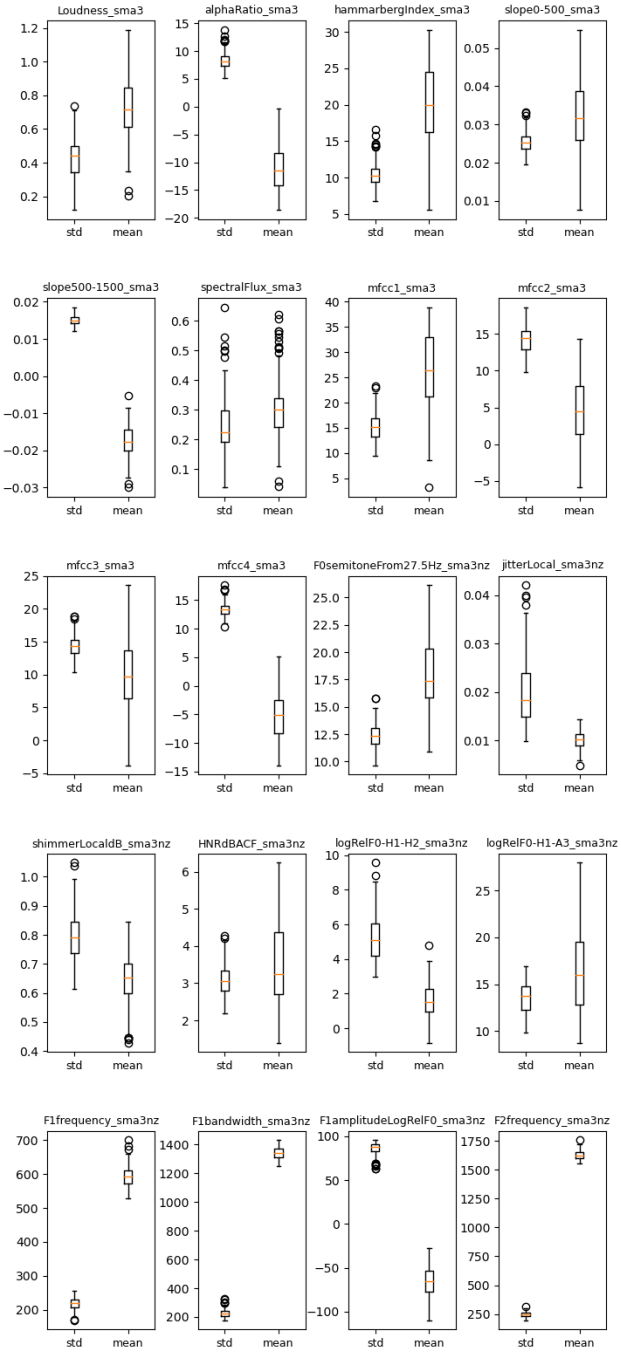


Figure 8.17: Distribution of the mean and std values per descriptor across all BAs

Description of attributes in terms of localized temporal information

Figure 8.17 illustrates the distribution of mean and standard deviation (std) values for each descriptor across attributes. While this figure presents a global behavior of all

attributes compared to each others, it distinctly reveals important variations between attributes in terms of encoded temporal information. To provide a closer look, let's consider the "Loudness_sma3" descriptor as an example. Notably, the mean and the std values exhibit substantial differences, ranging from 0.4 to 1.2 for the means and between 0.1 and 0.8 for the std.

8.6 Discussion and perspectives

In this chapter, we proposed a novel method that explains the nature of information encoded within binary-attribute-based speaker embeddings. This method, namely the three-world methodology, draws an interaction between the speech real world, the representation world and an informative world that presents any type of available information extracted from the real world. The idea is to find a mapping between the representation world and the informative world, which would provide more explanations about the representation world. The main goal of this chapter is to adopt this methodology to provide explanations of attributes through two levels. At the utterance-level, we provided acoustic and phonetic description of attributes. At the frame-level, we provided a phoneme and a temporal localization of attributes. The three-world methodology, along with the provided explanations of attributes, bring forth numerous noteworthy advantages and perspectives that we highlight in the following:

- *Flexibility*: The three-world methodology is flexible, straightforward to understand, easy to apply and practical. It represents a general concept that could be applied to any available or automatically extracted information about the speech data in the informative world. No additional annotations or manual labeling are required to explain information in the encoded representation. This explainability method operates akin to a self-supervised learning model; in this context, it can be referred to as a self-explanatory method. Also, it allows for flexible choice of the mapping function based on the requirements. From a broader perspective, it is essential to highlight that this method is not confined solely to binary attribute representation; instead, it has the flexibility to be applied to any discretized representation. In Appendix D §D.2.2, the step 3 methodology is applied to different binary representations in a different context.
- *Accurate and reliable description*: The achieved accuracy by attribute models in training indicates the capability of descriptors to effectively differentiate between the two classes of the attribute. The accuracy of these models underscores both the presence and robustness of the information carried by these descriptors in defining attributes. This performance significantly enhances confidence in the selection of the most influential descriptors. Additionally, the convergence of the attribute models and SLDA at the utterance-level description enhances trust in this description. Given that explainability lacks a well-defined formulation, we believe that validations using other mapping functions and evaluation with different ways are important to ensure reliable explainability of attributes.
- *Fidelity of attribute description*: Utilizing the BA-extractor train data to describe

attributes through attribute classifiers and SLDA proves advantageous, given the robust presence of encoded information. The attribute model has shown its fidelity and consistency to test data. As these attributes primarily represent neurons learned during the extractor training, we believe that applying the extractor to different data sets is not supposed to alter the information encoded in these attributes during training.

- *A dedicated mapping between attributes and input frames*: Thanks to the thresholding applied to the BA-extractor during training and the modification applied to access the frame-level information, we were able to map each attribute with its corresponding set of frames. This was possible through the activations of the MegaFrames corresponding to one attribute. The fact that only a subset of MegaFrames is active when an attribute is present was helpful and made the backpropagation into the DNN more straightforward, by focusing solely on highlighted MegaFrames. This added another dimension to the explainability of the attribute and provided a multi-level description.

- *Useful tool for phoneticians*: The obtained attribute descriptions serve not only as valuable aids for phoneticians in offering interpretations of speech representations, but it also may serve as powerful tools enabling them to uncover novel patterns and combinations of descriptors. For instance, the BA-vectors could be used in a classification task such as gender or emotion and then, by identifying the most contributing attributes to the classification permits to have a task-oriented explanation. The available description of these attributes would provide further insights into the phonetic information encoding specific vocal characteristics. An illustration of this example is provided in Appendix D.

Indeed, our proposed method achieved enhanced explainability level thus far. However, from a critical standpoint, it still lacks certain elements necessary to attain a complete level of interpretability that can be easily understood by a human. In the following, we highlight some of these elements:

- *Explainable but not fully interpretable*: While the obtained descriptions at both the utterance- and frame-levels provide insightful explanations, they remain somewhat challenging for individuals without expertise in phonetics and acoustics to fully comprehend. Consequently, the method lacks a higher level of interpretation tailored specifically for phoneticians. For instance, the variations between attributes in terms of phonetic descriptors and temporal information could be more easily explained by phoneticians in terms of higher-level vocal characteristics. Moreover, the exploration of phonemes and their classes uncovered an unexpected pattern within the stops class, underscoring its importance for specific attributes. This outcome deviates from typical findings in the literature and may benefit from interpretation by a phonetician. Another possibility is that it could be attributed to a potential hallucination within the BA-extractor or related to an error of the speech recognizer, as no ground-truth transcription is available.

- *Other mapping functions are to be explored*: Even though in this work we tested three mapping functions such as a machine learning function, a test statistic and backpropagation through the network, yet there exist many other mapping functions that could be applied. For instance, we could use information theory-based methods such as the mutual information or entropy between descriptors and attribute classes. Other mapping functions are tested for a different binary representation in §D.2.2.

- *Need for more data per attribute*: It is noteworthy that having larger sets S_0 and S_1 per attribute is important to provide explanations per attribute. The availability of more data per attribute boosts the stability of the attribute description and provides more trustful explanations. This would also help to explore further insights and to discover some composed explanations of the attribute.

- *Further analyses of temporal descriptions are to explore*: Indeed, the provided frame-level and temporal information descriptions are indicative and preliminary, and further studies are warranted. For instance, conducting phoneme and temporal descriptions directly on the training data would likely yield more reliable and robust results. This approach could provide a more general identification of certain attributes as effective detectors of phoneme classes. Furthermore, to enhance the robustness of descriptions and establish greater confidence in explanations, an additional exploration into aligning utterance-level phonetic descriptions with frame-level temporal information would be beneficial. Such an investigation would contribute to a more precise attribute description. Furthermore, it serves as a validation step for the mapping process between attributes and frame-level units.

Part III

Improvements and application of our approach in forensic context

Chapter 9

Calibration and application of BA-LR on forensically realistic database

9.1 Introduction 138
9.2 Global calibration 138
9.3 Selective fusion of attribute LLRs 140
9.3.1 Weighted fusion of attributes 140
9.3.2 Selection of attributes 141
9.4 Experimental protocol 141
9.4.1 Database description 141
9.4.2 Experiments setup 143
9.5 Calibration and speaker recognition performance 146
9.6 Discussion 149

Thus far, we have outlined the three steps of our approach designed to enhance the interpretability and explainability of a speaker recognition system. In this Chapter, our focus shifts towards the practical application of BA-LR framework, specifically involving steps 1 and 2, to report an interpretable LR for forensically realistic database. Due to data confidentiality concerns, step 3 could not be executed. In order to adapt BA-LR to the specific conditions of the forensic data, we apply a calibration method that aims to transform miscalibrated Log Likelihood Ratio LLRs into well-calibrated LLRs. This application provides also an opportunity to introduce a fusion and calibration approach for LLRs that improves both performance and calibration.

9.1 Introduction

During a visit to the Netherlands Forensic Institute (NFI), the author of this thesis applied BA-LR framework on one of their forensically realistic database. The goal is to further evaluate the generalization capability of BA-LR framework on other data. In this context, the mismatch in domain, conditions, and populations between training and evaluation may lead to poorly calibrated LLRs [269]. This miscalibration was previously observed and highlighted in Chapter 7 for certain evaluation datasets. In such scenarios, a calibration step becomes essential to transform these scores into well-calibrated LLRs. The traditional approach to calibration consists in employing an affine function with trainable parameters [127, 270]. These parameters are fine-tuned by optimizing an objective function. This function computes the LLRs of a development (Dev) set of target and non-target pairs, given their respective scores [271]. The learned parameters are used to transform the scores of the evaluation data (Test) into calibrated LLRs. The effectiveness of this transformation depends on how well the conditions of the Test pairs are reflected by the Dev pairs [272]. This is a common practice in speaker recognition evaluations organized by NIST [273]. *Logistic Regression* is the standard calibration approach frequently employed in speaker recognition [271, 273, 127, 270, 274, 119]. This method presents an affine transformation, shifting and scaling non-calibrated scores[1] to obtain well-calibrated LLRs.

In scenarios where multiple parallel scores are obtained using different ASpR systems, *Logistic Regression Fusion* is a frequently employed technique [275, 100, 273] that combines these scores, resulting in more accurate and well calibrated LLRs. Thus, this approach is not only beneficial for LLRs calibration, but also for performance improvement [127, 119, 272, 276].

In this chapter, we focus on applying BA-LR framework on forensically realistic data, namely NFI-FRIDA. This database has been introduced and evaluated only once in the literature [277, 278] using Vocalise software [103] based on DNN x-vector [47]. Through this application, we aim to achieve three main objectives: 1) Assess the generalization capability of BA-LR in a forensic context. 2) Improve the global calibration of the final LLRs. 3) Explore the impact of a Logistic Regression fusion of attribute LLRs on both, speaker discrimination and calibration performance. In the next sections, we start by defining the global calibration approach. Following that, we introduce a fusion approach of attribute LLRs. Subsequently, we detail the experimental protocol, including the NFI-FRIDA dataset as well as the applied methodology. Finally, we present and discuss the results in terms of speaker recognition performance and calibration.

9.2 Global calibration

As discussed in Chapter 7, in BA-LR framework, the attribute LLRs are computed using both the behavioral parameters of the attributes and the attribute values on both sides of the pair (see §7.4.1). It is important to recall here that the used behavioral

[1] We intentionally use the term "score" to refer to non-calibrated LLR

parameters are estimated from a reference population of the train dataset, namely the 6000 speakers of VoxCeleb2. In scenarios where an evaluation dataset presents different conditions and distinct quality, a mismatch between the training and evaluation datasets may arise, leading to non-calibrated LLRs. To address this mismatch, we train and apply a logistic regression model on the final LLRs. This model is mathematically defined as follows.

Let us consider a dataset $\{s_i, y_i\}_{i=1}^{n}$, where $S_i = (s_1, s_2, \ldots, s_m)$ is an m-dimensional variable, and the target variable Y_i is a binary variable, being 0 or 1. The logistic regression model is as follows:

$$\log(\frac{P(y_i = 1|s_i)}{1 - P(y_i = 1|s_i)}) = \alpha + \sum_{j=1}^{m} \beta_j \cdot s_{ij} \tag{9.1}$$

Where $P(y_i = 1|s_i)$ is the posterior probability of the positive class, given s_i, α represents the intercept. β_j is a regression coefficient and $\beta = (\beta_1, \ldots, \beta_m)^T$ is the regression coefficient vector. The logarithmic likelihood function is therefore expressed as follows:

$$l(\beta, \alpha) = \sum_{i=1}^{n} [y_i \cdot (\alpha + \sum_{j=1}^{m} \beta_j \cdot s_{ij}) - \log(1 + \exp(\alpha + \sum_{j=1}^{m} \beta_j \cdot s_{ij}))] \tag{9.2}$$

The global calibration of LLRs using logistic regression [272, 119, 100] is illustrated in Figure 9.1. Given a set of n comparison pairs $(X1_i, X2_i)$ where $i = 1...n$, S_i is a 1-dimensional variable representing the final $LLR_{X1_i,X2_i}$ scores, and Y_i is a target variable representing the ground truth of scores being target (i.e =1) or non-target (i.e. =0) pairs. The logistic regression model in this case is a univariate model with $m = 1$:

$$LLR'_{X_i,X2_i} = \alpha_G + \beta_G * LLR_{X1_i,X2_i} \tag{9.3}$$

Where α and β are scalars and denoted by α_G and β_G, respectively. The obtained $LLR'_{X1_i,X2_i}$ represent therefore the calibrated LLRs.

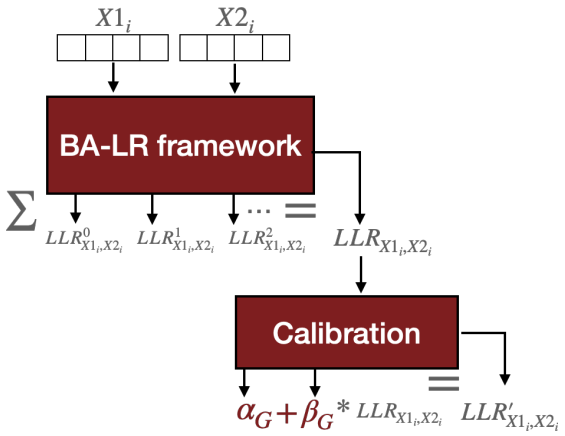


Figure 9.1: Global calibration of the final LLR using univariate logistic regression

9.3 Selective fusion of attribute LLRs

The primary objective of the fusion approach is to reduce the impact of the independence assumption between attributes in the final LLR computation [272]. In this section, we propose a weighted fusion of attribute-LLRs, rather than a straightforward summation of all attribute-LLRs. This approach selects only an effective subset of attributes. We pursue this method through two main concepts: 1) Using a multivariate logistic regression model for the weighted fusion of attribute LLRs. 2) Integrating a sparsity regularization in this model to retain only relevant attributes while discarding irrelevant ones.

9.3.1 Weighted fusion of attributes

The application of logistic regression fusion requires two key elements: 1) A training dataset consisting of scores from comparison pairs with known target and non-target ground-truth [279, 272]. 2) Multiple sets of scores from parallel comparison pairs conducted on the same data using different systems [100, 275, 276]. These systems could be different automatic systems. For instance, the work in [280] illustrates the fusion of LLRs from a forensic ASpR system with those derived from a forensic semi-ASpR system. It could be also systems that use different modelling techniques, or even different acoustic-phonetic systems where each system tackle information from a distinct phonetic unit within the same data [272].

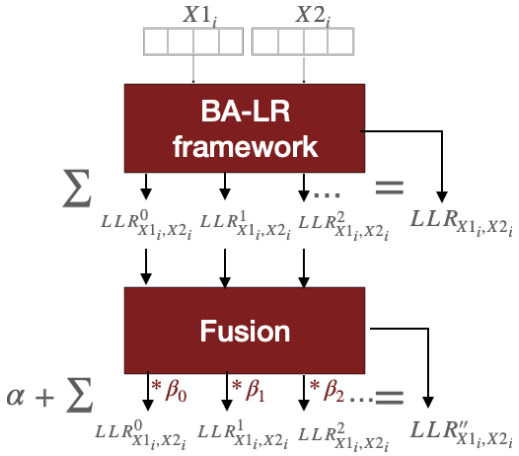


Figure 9.2: A weighted fusion of attribute-LLRs using a multivariate logistic regression model

In our case, each attribute represents a system that output an attribute LLR. We represent each comparison pair $(X1_i, X2_i)$ by an m-dimensional variable S, which comprises m attribute LLRs, denoted as $S_i = (LLR^0_{X1_i,X2_i}, LLR^1_{X1_i,X2_i}, ..., LLR^{m-1}_{X1_i,X2_i})$, and a binary target variable Y_i indicating whether the pair is a target or non-target. Logistic regression is then applied and modeled on multivariate data to optimise fusion

weights of the m attribute-LLRs in order to obtain a well-calibrated $LLR''_{X1_i,X2_i}$, as illustrated in Figure 9.2 and expressed in Equation (9.4).

$$LLR''_{X1_i,X2_i} = \alpha + \sum_{j=1}^{m} \beta_j * LLR^j_{X1_i,X2_i} \tag{9.4}$$

9.3.2 Selection of attributes

Regularizing logistic regression involves incorporating a term into the objective function, penalizing the distance from the estimated parameters to a default set of parameters. This technique is frequently employed to enhance the robustness of the calibration model and mitigate the risk of overfitting [281]. The most commonly used regularization are Lasso [282](L1 sparse regularization), ridge regularization (L2 regularization) [283] and Elastic-Net [284]. The use of *regularized logistic regression* for the calibration of speaker recognition systems has been studied in various works [270, 281, 285].

In our case, in order to push the logistic regression model to select only a truly relevant set of attribute-LLRs, a L1 regularization term [286] is added to the log-likelihood function of the logistic model as expressed in Equation.9.5. This regularization encourages *sparsity*, pushing the weights of some attributes to be exactly zero. This penalty term is usually added to the objective function to achieve the effect of sparsity and compression [287, 288, 77, 285]. The regularization parameter λ controls the strength of the penalty applied to the coefficients. By increasing the value of λ, the penalty on large coefficients becomes stronger. This encourages the model to shrink the coefficients towards zero, effectively reducing the impact of less important attribute scores and promoting sparsity in the coefficient vector. This L1 regularization results in a more interpretable and efficient model while helping to prevent overfitting [270].

$$(\hat{\beta}, \hat{\alpha}) = \underset{\beta,\alpha}{\operatorname{argmin}}\left(\frac{-l(\beta,\alpha)}{n} + \lambda \sum_{j=1}^{m} |\beta_j|\right) \tag{9.5}$$

9.4 Experimental protocol

This section outlines the experimental setup, providing details into the NFI-FRIDA database and offering an overview of the applied experimental protocol for the calibration and fusion approaches.

9.4.1 Database description

NFI-Forensically Realistic Inter-Device Audio (NFI-FRIDA) [277, 278] is a Dutch database comprising of speech samples recorded by various forensically relevant recording devices. This dataset comprises 302 male speakers representing a specific reference population. In the following, we describe the database in terms of recording devices and sessions.

Speakers

The dataset includes 302 male volunteer participants[2] who were not university educated. About 80% were aged between 18 and 35, with the remaining 20% aged up to 55. Among them, 50% had a native Dutch background, 25% had a Moroccan immigrant background, and another 25% had a Turkish immigrant background. Most participants were Amsterdam natives, and all recorded speech in the dataset is in Dutch, often including colloquialisms and street language.

Recording devices

The speech was simultaneously recorded with 5 devices in each session type. The devices were chosen to reflect conditions encountered in NFI casework. In this work, we are limited to three devices only, namely device 1, 4 and 5. We ignore devices 2 and 3 because they present similar quality as device 1. Devices 1 and 4 are recorded with a sampling rate of 48KHz, while device 5 has a sampling rate of 8KHz. The description of these devices is provided in Table 9.1 and is as follows:

Table 9.1: Recording devices description

	Recording device	**Session**
Device 1	Shure WH20 HQ Headset	1,2,3,4,5,6,7,8
Device 4	Shure SM58 far	1,2,3,4
Device 5	Intercepted telephone	1,2,3,4,5,6,7,8

- **Recording device 1 (device 1)**: A headset microphone that exhibits a high quality recording.
- **Recording device 4 (device 4)**: Recordings contain considerable reverberation and have a higher noise level. It presents low quality police interview recordings.
- **The intercepted telephone recordings (device 5)**: Provided by Dutch police. They are extracted through telephone interception system that is used in actual criminal investigations. Either an iPhone 4 or a Nokia 1280 telephone was used as shown in Table 9.2, according to the session. The iPhone 4 was chosen as it was the most widely used smartphone at the time of recording, and the Nokia 1280 was chosen to represent a cheap phone, often encountered in casework.

Sessions

Speakers were recorded across 16 sessions, engaging in spontaneous conversations on various topics over two days. These sessions were spaced apart by a minimum interval of one week. Each day comprised eight sessions recorded in diverse locations,

[2]The speakers are not from real forensic cases, but they represent a simulation of forensic cases

using different telephones, and varying in environmental noise, as detailed in Table 9.2. Each session lasted approximately 5 minutes, featuring telephone conversations between participants. In indoor sessions, a noisy environment included static radio noise, while outdoor sessions alternated between quiet and noisy street locations.

Table 9.2: Sessions description

Session	Location	Environment	Telephone
1	Inside	Silent	Nokia 1280
2	Inside	Silent	iPhone 4
3	Inside	Noisy	Nokia 1280
4	Inside	Noisy	iPhone 4
5	Outside	Calm	Nokia 1280
6	Outside	Calm	iPhone 4
7	Outside	Busy street	Nokia 1280
8	Outside	Busy street	iPhone 4

9.4.2 Experiments setup

In this section, we detail the data preparation step and provide specific information about the datasets used in each experiment. Following that, we present a comprehensive description of the experimental protocol, while precising the baseline employed. **It is noteworthy to mention here that all experiments are conducted using the second version of BA-LR, namely the Speech-based version.** This choice is motivated by the aim of presenting a logical and reasonable calculation of the LR in a forensic speech context.

Data preparation

In this experiment, we combine for the same device the data of each two sessions sharing the same location and environment in one session, as shown in Table 9.3. This table provides details on the number of utterances and speakers used in each combination of (device, session pair).

It is important to precise that, contrary to [277], no editing is applied to the data; all experiments use raw speech recordings under real conditions. However, given that the recordings of device 5 have a sampling rate of 8kHz, as they are telephone intercepts, we proceed with up-sampling these files to 16kHz. This adjustment is made to align with our BA-extractor.

Table 9.3: Experiment data description

Device	Sessions	#utterances	#Speakers
1	1&2	1,190	302
	5&6	1,186	302
	3&4	1,187	302
	7&8	1,184	302
4	1&2	1,190	302
	3&4	1,183	302
5	1&2	772	202
	5&6	766	203
	3&4	765	203
	7&8	768	204

Protocol description

To apply the calibration and fusion approaches on NFI-FRIDA data, we establish the protocol illustrated in Figure 9.3. For a given device$_i$-session$_j$, Dev and Test sets of utterances are selected and defined with 15-fold cross-validation. In each fold, utterances are randomly selected for the Dev and Test sets, ensuring that speakers are randomly assigned to each set with no overlap between speakers. For Dev and Test sets, the BA-vectors are firstly extracted using our BA-extractor trained on VoxCeleb2 dataset. Then target (tar) and non-target (non) comparison pairs are composed. The BA-LR framework is thus applied on these pairs to compute the attribute-LLRs and the global LLR. The Dev pairs are employed for training the logistic regression models, while the Test pairs are utilized for evaluation of SR performance and calibration. Details are provided in Figure 9.3 and are as follows:

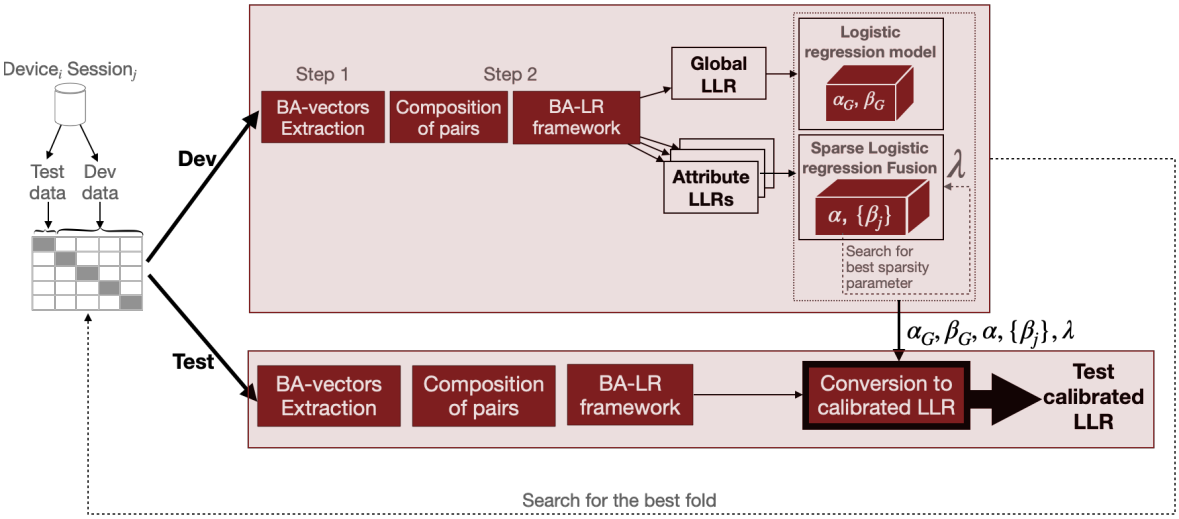


Figure 9.3: Description of experimental protocol using calibration and fusion approaches

- **Training phase**: In the global LLR calibration, the Dev global LLRs are employed to train the logistic regression model, determining the optimal shifting and scaling parameters, α_G and β_G. In the selective fusion, the Dev attribute-LLRs are firstly standardized, then used to train sparse logistic regression model, finding the intercept α and the optimal fusion coefficients of attribute-LLRs $\{\beta_j\}_{j=0}^{m-1}$. During the training of the latter, a grid search is conducted to identify the optimal sparsity parameter λ, ensuring the best discrimination performance and calibration on Dev set.

- **Testing phase**: In evaluation, we use the learned parameters α_G, β_G and we apply global calibration on LLRs as in Equation (9.3). For a more calibrated LLR and more accurate fusion of attribute-LLRs, we use the learned parameters α, β_j and λ as in Equation (9.4).

Table 9.4: Description of Dev and test sets of the best fold for each experiment

Device	Sessions	#Dev speakers	#Dev tar/non	#Test speakers	#Test tar/non
1	1&2	150	870/30,000	152	897/30,000
	5&6	150	872/30,000	152	884/30,000
	3&4	150	880/30,000	152	887/30,000
	7&8	150	866/30,000	152	884/30,000
4	1&2	150	870/30,000	152	897/30,000
	3&4	150	857/30,000	152	892/30,000
5	1&2	150	568/30,000	52	547/30,000
	5&6	150	540/30,000	53	555/30,000
	3&4	150	553/30,000	53	541/30,000
	7&8	150	553/30,000	54	546/30,000

We select, for both approaches, the same fold composition that provides the best average performance and calibration. The obtained results are later based on this fold composition. These experiments finally yield, for each approach, 4 models for device 1, 2 models for device 4 and 4 models for device 5. Details about the number of speakers and the target and non-target scores of Dev and Test sets are provided in Table 9.4. Note that the number of Non-target pairs is always restricted to 30,000 pairs.

Baseline

The experimental protocol employed in this work diverges from the protocol used in [277], which initially assessed NFI-FRIDA data for a speaker recognition task using the proprietary VOCALISE software. The unavailability of this software as open source, combined with the confidentiality of the database, prevents us from establishing a baseline for direct comparison with our results. For this reason, we establish our own baseline using the x-vector system employed in step 1 (§6.4.4). Please note that the results obtained with this baseline incorporate **all the Dev and Test comparison pairs**

for each experiment. Due to confidentiality constraints, we are unable to access the x-vectors to rerun these experiments only on test sets.

9.5 Calibration and speaker recognition performance

BA-LR generalisation ability

Table 9.5 presents the EER of the application of Speech-based version of BA-LR on Test sets, before and after fusion approach. Before applying the selective fusion, the overall ASpR performance of BA-LR in all experiments indicates its discrimination capability and its generalization[3] ability to the Dutch data. The superior discrimination performance observed on device 1 data, in contrast to device 4 and device 5, can be explained by the higher quality of recordings in device 1. Furthermore, device 4 and device 5 represent forensic conditions and telephone intercepts, respectively, which are not covered in either the training data of the BA-extractor or the attribute behavioral parameters used in the BA-LR framework.

Table 9.5: Speaker recognition performance of BA-LR Speech-based version on Test sets before and after selective fusion

Device	**Sessions**	**BA-LR**	**Selective Fusion**	
		EER (205 BAs)	**EER**	**#BAs**
1	1&2	1.037%	1.87%	132
	5&6	0.96%	1.2%	139
	3&4	1.22%	1.83%	149
	7&8	0.43%	0.5%	159
4	1&2	2.07%	2.37%	119
	3&4	4.27%	2.82%	144
5	1&2	10.05%	7.31%	101
	5&6	11.2%	7.84%	128
	3&4	10.72%	7.18%	127
	7&8	12.61%	7.59%	124

BA-LR Vs. baseline X-vector

For comparison reasons, results from the x-vector baseline on all comparison pairs are provided in Table 9.6. It is crucial to recall that the results presented with the baseline x-vector are computed using a combination of train and test comparison pairs. Any comparison made with our results against this baseline should be considered indicative only, as it reflects the average performance across both sets. Before fusion, Compared to baseline x-vector, an average slight absolute increase in EER of 0.85% for all devices is observed using BA-LR, except for device 4 of forensic conditions with 1.66% absolute increase of average EER.

[3]The training data predominantly consists of English speech samples.

Table 9.6: Speaker recognition performance of X-vectors on all comparison pairs. These results are indicative only, as they are calculated based on all comparison pairs corresponding to the combination of (device, sessions).

Device	**Sessions**	**X-vector**
1	1&2	1.02%
	5&6	0.85%
	3&4	0.74%
	7&8	0.28%
4	1&2	1.59%
	3&4	1.43%
5	1&2	8.16%
	5&6	9.53%
	3&4	9.7%
	7&8	11.1%

Table 9.7: $\text{Cllr}_{min/act}$ computed with BA-LR before (Non-Calibrated) and after (Calibrated) applying calibration and fusion approaches (results for the best fold)

Device-Sessions	**Non-Calibrated**		**Calibrated**			
			Global		**Fusion**	
	Cllr_{min}	Cllr_{act}	Cllr_{min}	Cllr_{act}	Cllr_{min}	Cllr_{act}
d1-1&2	0.04	0.60	0.04	0.08	0.07	0.10
d1-5&6	0.04	0.64	0.04	0.06	0.05	0.078
d1-3&4	0.04	0.64	0.04	0.06	0.07	0.08
d1-7&8	0.01	0.59	0.01	0.03	0.02	0.02
d4-1&2	0.08	1.71	0.08	0.10	0.10	0.10
d4-3&4	0.16	8.26	0.16	0.16	0.1	0.12
d5-1&2	0.35	8.78	0.36	0.38	0.26	0.30
d5-5&6	0.41	10.2	0.41	0.45	0.28	0.30
d5-3&4	0.35	10.0	0.35	0.38	0.26	0.27
d5-7&8	0.42	10.1	0.42	0.43	0.27	0.28

Calibration and fusion results

As shown in Table 9.7 the LLRs obtained with BA-LR are initially miscalibrated, which is particularly noticeable for device 4 and device 5 with an important difference between C_{llrmin} and C_{llract}. After calibration, the global calibration approach effectively converts these miscalibrated scores into well calibrated LLRs. As expected, this calibration is not supposed to impact the discrimination performance, and the EER remains unchanged. Interestingly, in addition to the calibration improvement, the selective fusion approach remarkably improves the overall discrimination performance of BA-LR on device 4 and device 5 (Table 9.5). Nevertheless, for device 1, where the recordings are of high quality, the fusion approach shows a slight increase in EER compared to the EER calculated using all BAs. This might be due to an overfitting of the model to the high-quality data in device 1. Moreover, it is noteworthy to highlight

that the BA-LR fusion approach outperforms the ASpR performance of the x-vector, especially on telephone intercepts of device 5, despite using a reduced number of binary attributes, as shown in Table 9.5, compared to the 256 floats of the x-vectors.

EER and C$_{llr}$ Vs. Number of selected attributes

Using the selective fusion model, each experiment results in the selection of a subset of attributes, as illustrated in Table 9.5. On average, the number of attributes selected for weighted fusion represents ∼67% of the initial total number of attributes in BA-LR (i.e., 205 BAs), for both device 4 and device 5 experiments. For more insights into this selection process, Figure 9.4 illustrates an example of the evolution of both EER and Cllr$_{cal}$ (i.e., C$_{llract}$-C$_{llrmin}$) with respect to increasing number of attributes selected by the λ of the sparse regularization, specifically for the optimal fold. As the values of λ decrease, and consequently, the number of attributes increases, the EER consistently decreases until reaching a certain number of attributes, after which it starts to rise again. The Cllr$_{cal}$ exhibits a parallel behavior to the EER, with the optimal EER aligning with the minimum Cllr$_{cal}$. This observation facilitates the identification of the optimal number of attributes that ensures both efficient discrimination and calibration performance.

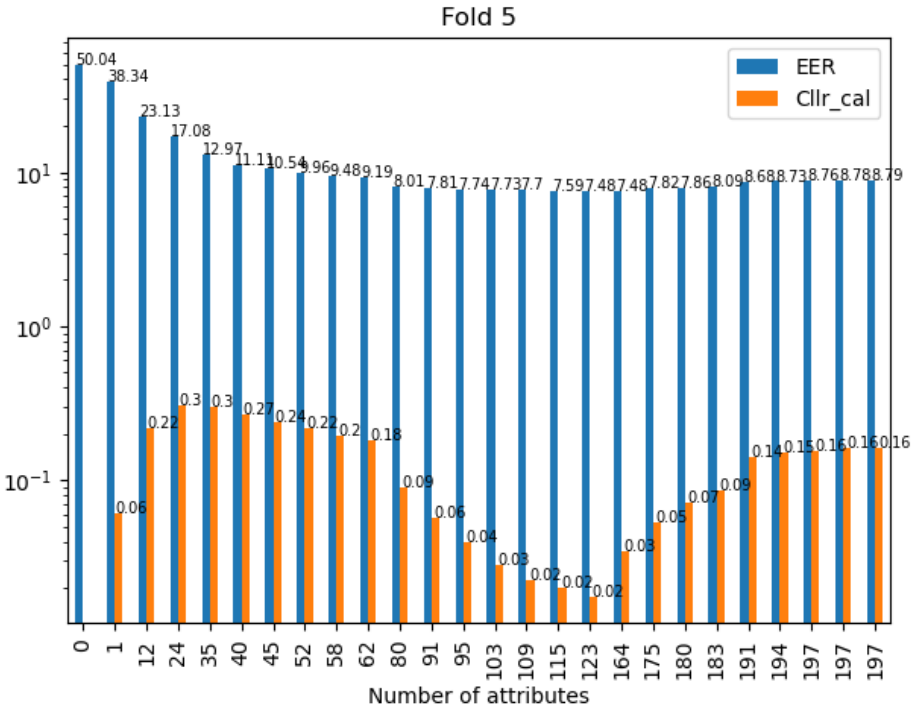


Figure 9.4: An example of EER and Cllr$_{cal}$ evolution using BA-LR, along with the number of attributes for the optimal fold

Given that these results are obtained using the Speech-based version of BA-LR, we show in Table E.2 and E.1 a comparison between the two BA-LR versions in terms of performance and calibration. Before fusion, the performance of both versions is comparable. After fusion and calibration, we obtain approximately same results for

both versions in terms of EER and $C_{llrmin/act}$.

9.6 Discussion

In this chapter, we evaluated the performance of BA-LR framework on the forensically realistic NFI-FRIDA database. A calibration and a fusion approaches were applied to address the mismatch between training and evaluation datasets. The two approaches use logistic regression model; The global calibration approach aims to produce well calibrated final LLRs, whereas the selective fusion approach aims to find an optimal fusion of attribute-LLRs using BA-LR to obtain more accurate and well calibrated LLRs. This fusion was found beneficial for BA-LR, as it assigns weights to relevant attribute-LLRs and entirely eliminates the influence of others, rather than simply summing all attribute-LLRs to determine the final LLR.

Even thought our BA-extractor was trained on VoxCeleb2, the attributes behavioral parameters are estimated on VoxCeleb2, and VoxCeleb2 being predominantly an English dataset (§6.4.1), the overall performance of BA-LR on NFI-FRIDA, a Dutch data, proves its strong generalization capability. Since this data reflects real forensic conditions, a mismatch is noticed in speaker recognition performance and calibration. This mismatch was more pronounced in device 5 comprising of telephone intercepts, as this specific data variability was not considered during the training of our BA-extractor. The global calibration approach is shown to address this mismatch by converting the LLRs into well calibrated LLRs.

Although the independence assumption is not explicitly imposed and respected in step 1 of our approach, the fusion approach allowed the regulation of the potential correlation between attributes by applying appropriate weights. By selecting an optimal subset of attributes only, the fusion approach effectively addressed the miscalibration of scores by enhancing not only the discrimination performance of BA-LR but also the calibration. Remarkably, it is even shown to outperform the discrimination of x-vector baseline in the case of device 5 experiments.

As a perspective, another possible solution to tackle the mismatch between train and evaluation data is to finetune the extractor on some telephone intercept samples and on dutch data recorded in forensic conditions. This was not possible in our case due to the confidentiality constraint on the NFI data.

From a broader perspective, the fusion approach applied to BA-LR scores reveals interesting interpretability aspect. By examining the weights assigned to the attribute-LLRs, one can gain direct insights into the importance of each attribute in the final LLR calculation. Combined with step 3 of our approach, the description of each attribute along with its weight permits to enhance the interpretability of the final LLR value.

However, while these results are indeed promising, it is crucial to approach the forensic application of BA-LR with caution. Further research is necessary for real-world deployment. Specifically, larger and more diverse databases should be experimented. This would help understanding the influence of the selected training database's sig-

nificance and the extent to which our findings can be generalized to specific cases commonly encountered in forensic contexts.

Chapter 10

Attribute-based binary auto-encoder

10.1 Introduction 152
10.2 SPINE: Sparse binarized speaker representation 152
10.2.1 Why sparsity? 153
10.2.2 SPINE model 153
10.3 BAE: attribute-based Binary Auto-Encoder 155
10.3.1 Binary auto-encoder model 155
10.3.2 Proposed attribute-oriented loss 156
10.4 Experimental protocol and analyses 158
10.4.1 Setup 159
10.4.2 Compliance with attribute-based criteria 160
10.5 Speaker recognition evaluation 162
10.5.1 Using cosine similarity scores 162
10.5.2 Application of BA-LR on BAE-vectors 163
10.6 Discussion and perspectives 165

Up to this point, the BA-extractor proposed in Step 1 has been utilized in all experiments. In this chapter, we introduce a new direction designed to enhance the performance of the binary-attribute-based representations of Step 1 of our approach. For this end, we explore two different solutions based on an auto-encoder able to produce binary embeddings.

10.1 Introduction

In the first step of our approach, we introduced an initial version of the binary-attribute-based extractor (BA-extractor). This extractor, although not optimal, meets the requirements for attribute-based representations, treating dimensions as binary attributes shared between speakers with relative independence. While this extractor proves beneficial for explainability, still it presents some limitations. Secondly, the objective of shared attribute modeling is indirectly and not explicitly driven or taken into account. Thirdly, binarization is not integrated into the modeling, but applied after the training of the model. Finally, its ASpR performance falls short compared to the baseline x-vector.

To address these limitations, we explore two alternative solutions for extracting binary and attribute-based vectors in an auto-encoder fashion, using ResNet x-vectors as input. The first auto-encoder, named as SParse Interpretable Neural Embeddings (SPINE), is initially introduced by [289] and applied in the context of JSALT workshop 2023[1] to extract sparse, interpretable and binary speaker vectors. The second extractor, termed Attribute-based Binary Auto-Encoder (BAE), is a novel approach proposed in this work with the specific purpose of directly generating binary and attribute-based representations. It is important to emphasize that the aim of the extractor is not only to extract binary vectors that improve ASpR performance, but also to ensure that these binary vectors meet the requirements of our goal representation. The adherence to these requirements is crucial for the subsequent applicability of BA-LR framework.

In this chapter, we begin by detailing the modeling process for the two extractors, SPINE and BAE, respectively. We then outline the experimental protocol, covering model configurations and additional analyses to verify the compatibility of binary vectors with the BA-LR framework. Subsequently, we assess the performance of these vectors in terms of speaker recognition, followed by an application of BA-LR on the compatible binary vectors. Finally, we summarize the chapter and engage in discussions and perspectives.

10.2 SPINE: Sparse binarized speaker representation

During the 2023 edition of the JSALT workshop[2], the topic of explainability in the case of speaker diarization system was tackled[3], aiming to answer the question, "who speaks? when? and why?". The ultimate goal of the project[4] is to explore speaker embeddings to provide explainable automatic diarization results.

Inspired from our three-step approach, the proposed speaker model involves learning a representation space specifically designed for interpretability [290] . This extractor[5]

[1]The Jelinek Summer Workshop on Speech and Language Technology 2023

[2]This work was supported by Johns Hopkins University and H2020-MSCA ESPERANTO

[3]We participated in a collaborative effort within an international, multidisciplinary team.

[4]https://jsalt2023.univ-lemans.fr/en/explainability-for-diarization.html

[5]No publication about this work is available for the moment

is mainly based on the work in [289] that generates interpretable word representations. The resulting representations are referred to as SParse Interpretable Neural Embeddings (SPINE). This extractor aims to promote two desirable properties in the representations: *sparsity* and *non-negativity* [291, 292, 293]. In this section, we justify the choice of these two properties then we describe SPINE architecture used in this step.

10.2.1 Why sparsity?

The idea of introducing sparsity in DNN training is mainly inspired from our brain functioning. It is shown in many studies such as [294] that our biological neurons are silent most of the time, with the exception of a percentage of neurons getting activated all at the same time. It is shown that given a specific stimulus, only a highly selective, small subset of neurons will activate. Sparsity is a property used frequently in DNN models to push their explainability and interpretability [295, 296]. The goal is that the sparsity constraint decorrelates the overall information distributed over the representation by focusing only on a set of neurons being activated for a specific output. Thus, the contribution of any neuron to the explanation of the input data can be easily determined [295, 297]. Several recent works, mainly in the field of natural language processing, employed this property to enhance interpretability in word embeddings [292, 298, 289], language models [296] and extraction of semantic properties [299]. These studies indicate that sparsity pushes the explainability aspect of the DNN models [289, 297]. Another key advantage of sparsity is that it is independent of the DNN architecture, it can be rather incorporated into any DNN structure.

10.2.2 SPINE model

SPINE [289, 298] is a k-sparse auto-encoder that imposes constraints on the latent space, ensuring that only k neurons are active at any given time [300]. Figure.10.1 illustrates the architecture of SPINE auto-encoder.

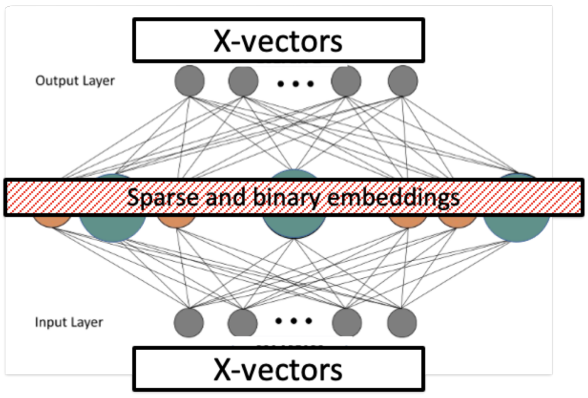


Figure 10.1: SPINE architecture adapted to speech

It is composed of an encoder with a single linear layer followed by a HardTanh activation function [290], expressed in Equation (10.1). This encoder takes x-vectors (X) of 256 dimensions as input and produces sparse embeddings Z^X of 500 dimensions in the latent space. In contrast to conventional auto-encoders, SPINE diverges by not compressing the representation within the latent space; instead, it expands the number of dimensions. This divergence is motivated by the fact that sparsity inherently achieves information compression. Subsequently, a decoder, consisting of a linear layer, reconstructs the input X from these sparse embeddings.

$$\text{HardTanh}(x) = \begin{cases} 1 & \text{if } x \geq 1 \\ 0 & \text{if } x \leq 0 \\ x & \text{Otherwise} \end{cases} \tag{10.1}$$

In order to impose sparsity constraint into the auto-encoder, the following loss function is being minimized during training:

$$L(D) = RL(D) + \lambda_1 \cdot ASL(D) + \lambda_2 \cdot PSL(D) \tag{10.2}$$

where D denotes the set of input x-vectors. RL is the reconstruction loss as expressed in Eq.(10.3). ASL is the average sparsity loss and PSL is the partial sparsity loss that enforce k sparse activations in the latent space H [289]. λ_1 and λ_2 are the weights given to ASL and PSL losses, respectively.

$$RL(D) = \frac{1}{|D|} \sum_{x \in D} ||X - \hat{X}||_2^2 \tag{10.3}$$

The goal of the ASL loss, expressed in Equation.(10.4), is to penalize any deviation of the observed average activation value $\rho_{h,D}$ from the desired average activation value $\rho^*_{h,D}$ of a given neuron, over a given data set.

$$ASL(D) = \sum_{h \in H} (\max(0, \rho_{h,D} - \rho^*_{h,D}))^2 \tag{10.4}$$

The PSL loss, as expressed in Eq.10.5, serves to penalize values that are neither close to 0 nor 1, pushing them close to either 0 or 1.

$$PSL(D) = \frac{1}{|D|} \sum_{x \in D} \sum_{h \in H} (Z^X_h \cdot (1 - Z^X_h)) \tag{10.5}$$

After the training phase, SPINE sparse embeddings are extracted from the latent space. Binary vectors are therefore obtained by thresholding, preserving zero values while converting others to 1.

However, as outlined in Chapter 6 regarding the BA-extractor, the post-training binarization process is not the most effective solution. This method tends to set all activations, both small and high, to 1. A more favorable approach would be to train a dedicated binary extractor explicitly designed to produce binary vectors. Moreover, while the auto-encoder does promote sparsity during training for reconstruction purposes, it lacks constraints that encourage dimensions to exhibit shared patterns among speakers.

10.3 BAE: attribute-based Binary Auto-Encoder

Drawing inspiration from the architecture and the constraints incorporated in SPINE, we propose a more purpose-oriented auto-encoder, referred to as attribute-based Binary Auto-Encoder (BAE). In the literature, most of the works that use binary auto-encoders are mainly for hashing purposes [301, 302], for preserving information [303, 304] or for data compression [305, 306, 307]. More aligned with our work, [308] uses a textual auto-encoder with latent space consisting of binary vectors for a text modelling task. It is noteworthy that the training of a binary auto-encoder is not straightforward due the non-differentiability of the gradient during training. The generation of binary vectors in the latent space makes the backpropagation of the gradient not possible. In this context, we aim to design a BAE model that meets the following criteria:

- To push dimensions to exhibit attribute-like behavior, shared across groups of speakers.
- To directly generate binary embeddings.
- To address the non-differentiability of gradient during the training in the case of binary auto-encoder.

In the following, we describe the architecture of the binary auto-encoder model and introduce the proposed loss functions designed to guide the model in generating the desired representations.

10.3.1 Binary auto-encoder model

Before delving into the proposed architecture of the BAE, we define a commonly used technique, namely Straight-through estimator, that makes the training of the BAE possible.

Straight-through estimator

The Straight-Through Estimator (STE), as introduced by [309], serves as a technique for training neural networks involving discrete latent variables, such as binary codes or discrete embeddings. It addresses the inherent challenges associated with back-propagation through non-differentiable operations. In the typical neural network back-propagation process, gradients flow backward to update model parameters. However, with discrete binary variables, these gradients are often undefined due to their non-differentiability. STE tackles this issue by employing a "relaxed" or "straight-through" gradient during the backward pass. This approach treats the discrete variable as if it is continuous, facilitating the flow of gradients as if the variable is non-discrete. The gradients are approximated using this continuous relaxation, providing a means to update model parameters. STE was used in various works such as [308, 310, 311, 305].

Architecture

The proposed BAE takes the baseline 256-dimensional x-vectors as input with the objective of reconstructing them. The BAE architecture, illustrated in Figure.10.2, consists of an encoder, which includes two linear layers with 256 and 512 units, followed by ReLU+batch normalization and Tanh activations, respectively. A representation of 512 dimensions, denoted as z, is obtained in the latent space. In the forward pass, this representation is binarized with a thresholding function that converts negative values or zeros to zeros and values strictly superior to zero to ones. This binary representation is therefore denoted as the *BAE-vector*. The decoder is composed of a linear layer of 512 followed by a Tanh activation and a linear layer of 256 units. In the backward pass, the gradient should back-propagate without passing through the BAE-vector. The Hardtanh function (Equation.10.1) is applied to clamp the gradient between -1 and 1.

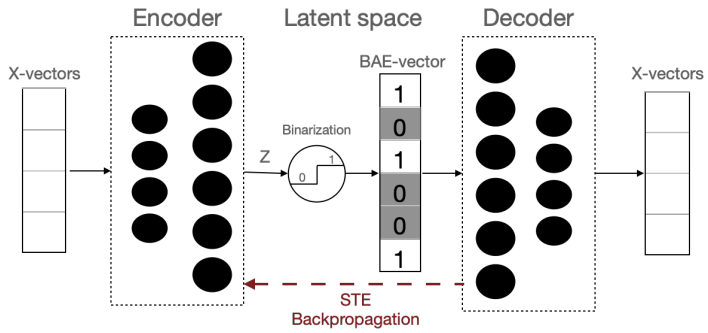


Figure 10.2: Architecture of the attribute-based Binary Auto-Encoder

The whole model is trained with two objective functions. In addition to the conventional reconstruction loss (MSE) in auto-encoders that takes as input x-vectors and tries to reconstruct them, we introduce a sparsity loss described in the next section.

10.3.2 Proposed attribute-oriented loss

The goal of the attribute-oriented loss is to push the model to encode binary representations modeled by shared discriminant attributes between speakers. In order to clarify the idea of this loss, let's consider in Figure 10.3 a toy example illustrating the relationship between utterances activations before binarization and speaker profile. It is important to remind that an attribute is considered present in the profile, if the sum of all utterance activations per attribute for a speaker is non-zero (refer to Chapter 7). We remind also that typicality is linked to the profile, computed as the presence frequency of attribute among speakers profiles. So back to our BAE, **what if we regulate the activations of the latent space dimensions before binarization to ensure that only a subset of speakers has a particular dimension present in their profiles?**

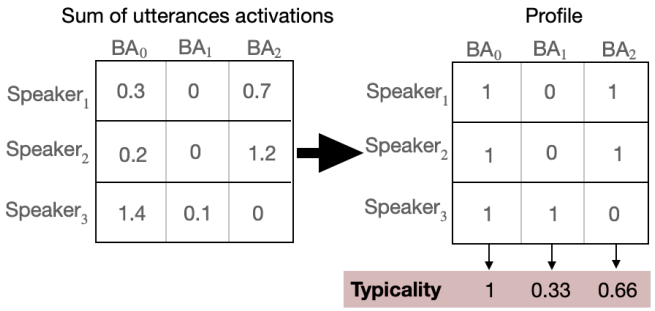


Figure 10.3: A toy example illustrating the relationship between utterances activations and speakers profiles used to compute the typicality

To address this, we propose a loss that aims to guide the dimensions towards achieving a desired presence frequency among speakers. This encapsulates the concept of typicality, where an attribute may be rare, moderately present, or typical among speakers. Consequently, ensuring the absence of an attribute in the speaker's profile entails driving the sum of activations across all their utterance vectors to 0. Conversely, for an attribute to be considered present in the speaker's profile, it suffices for only one vector of the speaker to possess this attribute.

The formulation of this loss is mainly inspired from the ASL loss of SPINE [289, 300] in Equation 10.4, while applying some modifications to adapt it to our need. More precisely, instead of constraining k-sparsity in the vectors of the latent space, we impose more strict constraint that pushes each dimension to follow a specific sparsity while considering speakers. To do so, a specific organization of the input batch of x-vectors is followed in this work.

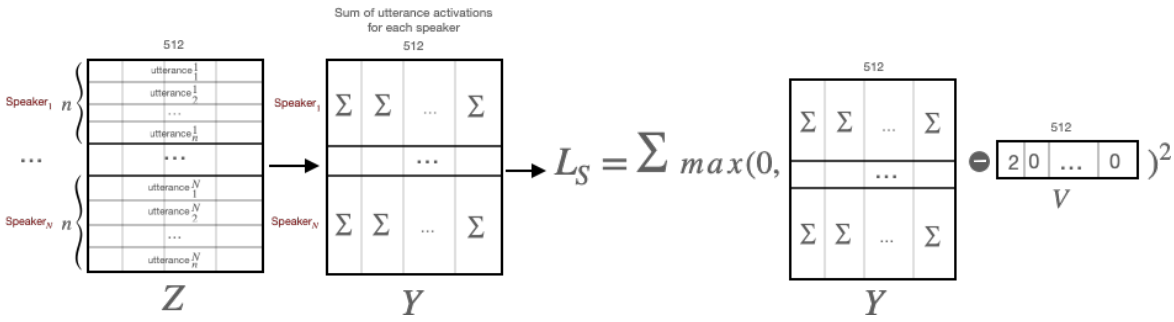


Figure 10.4: Illustration of the sparsity loss computation during training using the latent space representation before binarization

Figure.10.4 illustrates the batch organization in the latent space and the calculation of the proposed sparsity loss, following the annotations defined below:

- Let X be the input batch of 256-dimensional x-vectors. This batch is structured into sets of N speakers, where each speaker package consists of n x-vectors belonging to that particular speaker. X is therefore a matrix of size $(N * n, 256)$.

- Let Z be the output batch representation from the encoder of 512 dimensions, constituting the latent space. Instead of compressing the representation of the latent space, we opt here to increase the dimensionality due to the sparsity aspect of vectors. These representations are the Tanh activations ranging between -1 and 1. Z is a matrix of size $(N * n, 512)$.

- Let Y be the obtained speakers summary batch representation. This batch is calculated following equation.10.6 by summing all n activations of each speaker in each dimension. Y is a matrix of size $(N, 512)$.

$$Y_{i,j} = \left(\sum_{k=1}^{n} Z_{k,j} \right) \tag{10.6}$$

- Let V be a vector of 512 dimensions representing the desired frequency presence of attribute. This vector is generated randomly with values between 0 and n strictly.

As shown in Figure.10.4, given Y and V, the final attribute-oriented loss is calculated in such a way that the sum of n activations of each speaker in one dimension should be as close as possible to the desired frequency of that dimension in V. Since the dimensions activations are between -1 and 1, then the sum of the values would be always less than n. This sum for each speaker is subtracted by the desired dimension frequency of V. The result is then subjected to a max operation between 0 and the computed value to avoid negative values, followed by squaring. The summation of all these values represents the final loss, as expressed in Equation.(10.7) for one batch during training.

$$L_S = \sum_{i} (\max(0, Y_{i,j} - V_j))^2 \tag{10.7}$$

The MSE loss in this case is expressed as follows:

$$MSE = \frac{1}{N * n} \sum_{i=1}^{N*n} (x_i - \hat{x_i})^2 \tag{10.8}$$

The total loss to be minimized during training is therefore expressed in Equation.(10.9), where λ is the weight given to the attribute-oriented loss.

$$L = MSE + \lambda \cdot L_S \tag{10.9}$$

10.4 Experimental protocol and analyses

In this section, we first setup our experiments, along with models configurations. Next, we verify the compatibility of SPINE and BAE binary vectors with BA-LR framework.

10.4.1 Setup

In the following section, we present some details concerning the setting and training of the two models. Specifically, for the SPINE system, we specify the weights assigned to the sparsity loss functions. Additional details regarding the training parameters of the SPINE model will be provided in an official publication of the work conducted in JSALT2023[6] by the authors. We also detail the training of the BAE model, showing the evolution of the two losses.

SPINE model

The sparsity of SPINE vectors is mainly controlled by the two weights λ_1 and λ_2. Based on the sparsity level, three systems are proposed, as depicted in Table.10.1. It is important to note that sparsity indicates the percentage of zeros. All three systems are trained using VoxCeleb2 x-vectors extracted with our ResNet baseline explained in Chapter 6. The evaluation of these systems as well as the baseline is performed on VoxCeleb1[7] trials.

Table 10.1: SPINE configuration for three systems

System	λ_1	λ_2	Sparsity
SPINE-15%	1	1	15%
SPINE-50%	1	10	50%
SPINE-70%	50	10	$\sim$70%

BAE model

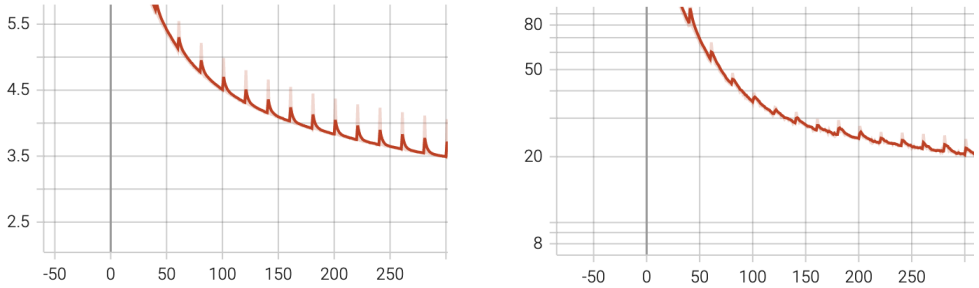
In this experiment, we setup our BAE model as follows:

- **Batch configuration**: N=27 speakers and n=10 x-vectors per speaker for each batch, which yields to a batch of 270 x-vectors.
- **Training parameters**: The number of epochs=2000, learning rate=0.001, λ=0.01.
- **Train and test data**: The x-vectors of VoxCeleb2 and VoxCeleb1, respectively, extracted with the baseline ResNet.
- **Evaluation protocol**: We employ the same experimental protocol for composing comparison pairs, as described in Table.7.1.

Figure 10.5 illustrates the evolution of MSE and attribute-oriented losses over a specific number of training epochs of BAE model. Notably, the attribute-oriented loss

[6]https://jsalt2023.univ-lemans.fr/en/explainability-for-diarization.html
[7]https://www.robots.ox.ac.uk/ vgg/data/voxceleb/meta/veri_test.txt

exhibits a more rapid decrease compared to the MSE loss, justifying our choice of its smaller weighting factor λ in relation to MSE. Additionally, a closer examination of the evolution of these two losses reveals a reciprocal relationship: as the MSE decreases, there is a slight increase in attribute-oriented loss, and vice versa. This behavior can be associated to the distinct optimization directions of the two losses. Throughout training, the model's goal is to find an optimal point that simultaneously satisfies and minimizes both objectives. The best optimal model is chosen based on the EER found for VoxCeleb1.



(a) MSE loss (b) Attribute-oriented loss

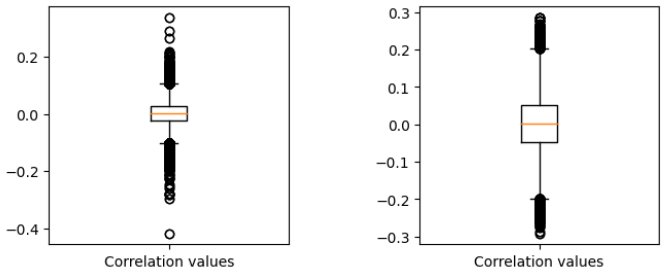
Figure 10.5: Snapshot of the losses evolution during the training ofBAE model

10.4.2 Compliance with attribute-based criteria

To apply the BA-LR framework on binary vectors, it is crucial for these vectors to conform to the predefined criteria of attribute-based representations. These criteria primarily involve ensuring the decorrelation or independence between dimensions and promoting attribute-like behavior, treating dimensions as attributes shared between subsets of speakers. In the following, we focus on the verification of these two aspects in binary vectors.

Dimension correlation

To verify the decorrelation assumption among the dimensions of the BAE-vector as well as the SPINE-vector, we compute the Pearson correlation. For SPINE-vectors, we opt to work with the best performing system, SPINE-15%.



(a) BAE-vectors (b) SPINE-15%-vectors

Figure 10.6: Distributions of Pearson correlation values between dimensions

Figure 10.6 depicts the distribution of correlation values of the two vectors, revealing consistently low values, falling within the range of -0.3 to 0.3. Notably, this correlation range closely resembles that of the BA-vectors.

Attribute-like behavior

The attribute-like behavior of a dimension is indirectly related to its typicality among speakers. To verify this behavior in binary vectors, we calculate the typicality values of dimensions across the train dataset for each binary vector.

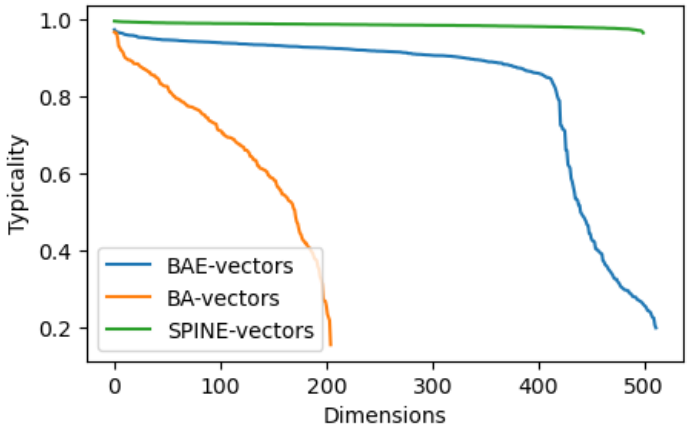


Figure 10.7: Sorted typicality values across dimensions of the three binary vectors.

Figure 10.7 shows the typicality values corresponding to the dimensions of BA-vectors, SPINE vectors and BAE-vectors. In the case of BAE-vectors, the initial 400 dimensions consistently display high typicality, ranging between 1 and 0.8. Conversely, the final 100 dimensions exhibit lower typicality values (i.e., 0.8-0.2). In contrast, BA-vectors exhibit a more uniform distribution of typicality values across all 205 dimensions. It is clear that contrary to BA-vectors and BAE-vectors, the dimensions of SPINE vectors do not exhibit attribute-like behavior. All dimensions in SPINE-vector consistently demonstrate the same behavior across all speakers, always present and

highly typical. This violates a key requirement of the binary-attribute-based representation (§.6.3.1), which specifies that attributes should be shared among groups of speakers.

Given this behavior, where SPINE dimensions do not adhere to the characteristics of attributes, these vectors are only binary vectors and not binary-attribute-based vectors. **Applying the BA-LR framework of Step 2 in our approach to SPINE-vectors is therefore unfeasible**. However, this does not prevent the application of Step 3 on SPINE vectors to explore and explain information encoded within these vectors. This study is left in Appendix D.

10.5 Speaker recognition evaluation

In this section, we conduct an evaluation in terms of speaker recognition performance using two different scoring systems. 1) Using cosine similarity to evaluate for all systems. 2) Through the application of BA-LR scoring for BAE system exclusively.

10.5.1 Using cosine similarity scores

Table 10.2: Speaker recognition performance of the three systems on VoxCeleb1 in terms of EER using cosine similarity scoring

Systems	**BA-extractor**	**BAE**	**SPINE**		
Binary vectors	**BA-vectors**	**BAE-vectors**	**SPINE-15%**	**SPINE-50%**	**SPINE-70%**
#Dimensions	205	512	500	500	500
Average Sparsity level	65%	75%	15%	50%	70%
EER[1]	3.42%	2.22%	1.66%	2.6%	3.3%

[1] The baseline x-vector exhibit an EER of 1.37%.

Table 10.2 illustrates the ASpR performance in terms of EER based on the five different binary representations, namely BA-vectors,BAE-vectors and the three variants of SPINE-vectors[8]. The EER is calculated based on the cosine similarity scores. The overall ASpR performance of these systems highlights the efficacy of the binary representations in discriminating between speakers. Notably, BAE-vectors demonstrate a notable reduction of 1.2% in EER, compared to the BA-vectors (refer to Chapter 6). This improved performance is thought to be comparable to the baseline x-vector, taking into account the binary aspect and dimensionality. On the other hand, the best SPINE-vectors, namely SPINE-15%, demonstrate superior performance compared to both BA-vectors and BAE-vectors, with a decrease in EER of approximately 0.29% compared to the baseline x-vector. However, it's essential to consider this performance in conjunction with the average sparsity level of these vectors relative to others. This trade-off between ASpR performance and sparsity is illustrated in the table for the

[8]More evaluation results are provided in §.D.2.1

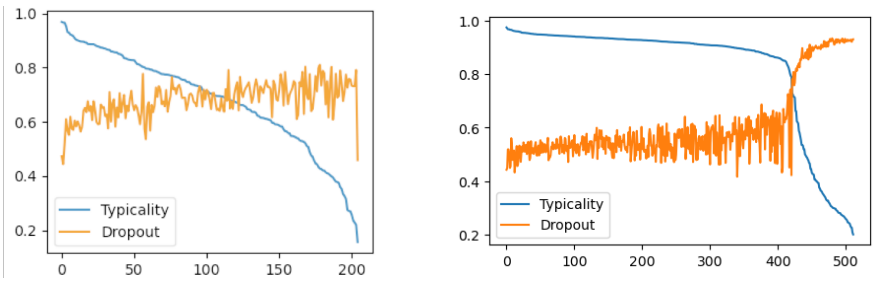
three variants of SPINE-vectors, namely SPINE-15%, SPINE-50%, and SPINE-70%. Notably, the decrease in ASpR performance becomes more pronounced with higher sparsity levels.

10.5.2 Application of BA-LR on BAE-vectors

In this section, we begin by estimating behavioral parameters such as Typicality, Drop-out and Drop-in using BAE-vectors. The ASpR performance of BAE-vectors is then evaluated using the two versions of BA-LR in a ASpR task.

Behavioral parameters analyses

Behavioral parameters such as Typicality and Drop-out are estimated on the train set of BAE-vectors following Equations (7.1) and (7.4), respectively. Figure 10.8 illustrates the relationship between typicality and drop-out values of BAE-vectors in comparison with BA-vectors. In contrast to BA-vectors, many dimensions in BAE-vectors exhibit high typicality values. Notably, the first 400 attributes have typicality values ranging from 1 to 0.8, while only 112 attributes fall within the range [0.8, 0.2]. Drop-out values are observed in the range [0.4, 0.7] for attributes with high typicality, whereas they increase in the range [0.7, 0.9] for attributes with low typicality. This behavior of drop-out was expected. This reflects the idea that when an attribute is present in the profile while its occurrence is observed in very few utterances, this yields to a high drop-out.



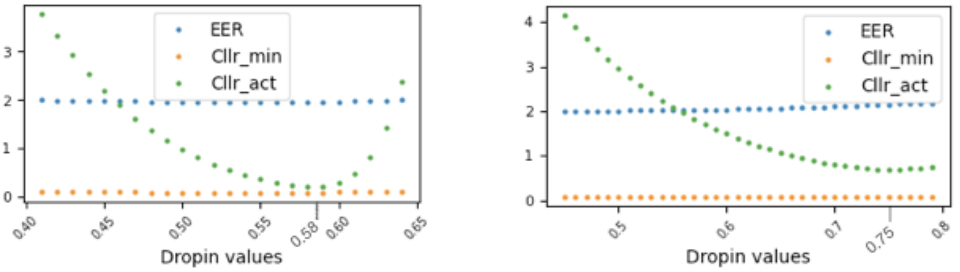
(a) 205-dimensional BA-vectors

(b) 512-dimensional BAE-vectors

Figure 10.8: Relationship between typicality and drop-out of BA-vectors and BAE vectors

Drop-in parameter on the other hand is estimated while setting multiple values of *Din* then choose the optimal value that finds the minimum difference between Cllr_{act} and Cllr_{min} for some trials of the train data, as described in Table.7.1. The evaluation of these trials is performed using the two versions of BA-LR. Thus, a *Din* factor should be find for each version. Figure.10.9 shows the search process for the optimal value of *Din* for BAE-vectors corresponding to each BA-LR version. While the drop-in is assumed to quantify the noise in the data, the values 0.58 and 0.75 for DNA-inspired

and Speech-based, respectively are considered high Din values. This is to be compared with BA-vectors, where Din is 0.12 and 0.26 for DNA-inspired and Speech-based, respectively (See Figure.7.7).



(a) Using DNA-inspired, $Din = 0.58$

(b) Using Speech-based, $Din = 0.75$

Figure 10.9: Search for the optimal drop-in value for each version of BA-LR using the train set of BAE vectors

Speaker recognition performance

Table 10.3 provides an overview of the evaluation results for the BAE system with BA-LR scoring in a speaker recognition task. This evaluation integrates also a trace of the information loss across different phases of the BAE system, including the input x-vectors X, the sparse latent space vectors Z, the binary BAE-vectors, and the reconstructed x-vectors $\widehat{X}$. This is to firstly quantify how well the auto-encoder preserves input information during reconstruction.

Table 10.3: Speaker recognition performance of BAE system and BA-extractor on VoxCeleb1 in terms of EER and $Cllr_{min/act}$

	Input	**Latent space**			**Output**	**BA-extractor**	
	X	**Z**	**BAE-vectors**		**X̂**	**BA-vectors**	
#Dimensions	256	512	512		256	205	
Evaluation	Cosine	Cosine	Cosine	DNA-inspired	Speech-based	Cosine	DNA-inspired	Speech-based
EER	1.37%	1.75%	2.22%	1.96%	2.46%	1.80%	3.7%	3.5%
Cllr$_{min/act}$	0.057/0.81	0.07/0.91	0.073/0.83	0.08/0.15	0.097/0.58	0.073/0.83	0.14/0.31	0.13/0.48

In the latent space, the obtained sparse vectors Z exhibit a marginal absolute increase in EER of only 0.38% compared to the input. Transitioning to the binary version of these vectors results in a further absolute EER increase of 0.47%, using cosine similarity scores. In terms of reconstruction, the ASpR performance of $\widehat{X}$ vectors indicates a low information loss compared to the input, with an absolute increase in EER of approximately $\sim 0.4\%$.

Applying the two versions of BA-LR on BAE-vectors reveals interesting discrimination performance with an average absolute loss of $\sim 0.84\%$ compared to the input. Remarkably, in comparison to BA-vectors initially proposed in Step 1, the BAE-vectors

achieve a significant absolute decrease, averaging around $\sim$ 1.39% in EER. This improvement is more pronounced using DNA-inspired version than Speech-based version of BA-LR. However, in terms of Cllr, it is important to note also that BA-LR scores are not well calibrated, especially for Speech-based.

10.6 Discussion and perspectives

In this chapter, we introduced two auto-encoders designed to generate binary and attribute-based representations. The first extractor, namely SPINE, is motivated by incorporating sparsity in the latent space to facilitate binarization. This auto-encoder aims to encourage sparsity, which yield into more interpretable representations. However, utilizing this extractor necessitates an extra binarization step on the produced sparse vectors, and the attribute-based behavior is not inherently modeled in the representations. The second extractor, namely BAE, is a binary auto-encoder proposed in this work to directly generate binary and attribute-based representations. This extractor imposes constraints on the latent space to guide the binary vectors towards the desired representation. This is accomplished through the introduction of an attribute-oriented loss function that pushes dimensions to exhibit attribute-like behavior concerning speakers.

The overall performance of SPINE and BAE binary vectors in ASpR task evaluated on VoxCeleb1 using cosine similarity is highly promising. Notably, they significantly outperform the performance of the initially proposed BA-vectors. Compared to BA-vectors, binary vectors generated by the best SPINE system exhibited an absolute reduction of 1.76% in EER, whereas BAE-vectors showed an absolute decrease of 1.2% in EER.

In contrary to BAE-vectors, SPINE-vectors are shown to be incompatible with the application of the BA-LR scoring because of the absence of attribute-like behavior in the representations. This reinforces the idea that not every binary vector is necessarily an attribute-based vector. Without explicitly emphasizing this aspect during the training of the auto-encoder, this behavior is not straightforward.

The proposed BAE-vectors demonstrated significantly improved results when using the BA-LR scoring on VoxCeleb1, surpassing the performance of the BA-extractor. Notably, it achieved a noteworthy reduction in EER, with an average decrease of approximately $\sim$1.39% using both versions of the BA-LR framework. These results of the proposed binary auto-encoder are very encouraging and promising, providing further evidence of the high potential of our three-steps approach in terms of improved performance. While the performance of the BAE is limited by the quality of the input x-vectors, employing more accurate input vectors would undoubtedly enhance the overall performance of the binary vectors.

However, still the proposed binary auto-encoder needs further exploration and improvements. First, even though the overall training of the auto-encoder aims to vary attribute typicality, choosing the right convergence point for the model is crucial to avoid exaggeration and reduce the reliability of attributes. Both drop-out and drop-in

are impacted by this behavior. Drop-out reflects the idea that when an attribute is rare among speakers, its probability of error becomes higher. It appears that when an attribute is present in the profile while its occurrence is observed in very few utterances, this yields to a high drop-out. This is likely due to the imposed constraint and the training process of the auto-encoder, where each batch contains only 10 samples of the speaker. One potential solution could be to consider all samples of each speaker in a batch. Drop-in is also shown to be very high, indicating that the encoding of binary vectors is noisy. These two behavioral parameters impact the attribute LLRs interpretability.

Second, considering that the optimal model of the BAE is picked up looking at the minimal EER on VoxCeleb1, it is important to acknowledge that this may not necessarily represent the truly optimal model. Therefore, two considerations arise: firstly, evaluating this model on other datasets, and secondly, exploring alternative criteria for selecting the optimal model, such as basing the choice on the estimation of behavioral parameters from the training data.

Third, given that the results of BA-LR in terms of $Cllr_{min/act}$ reflect miscalibrated LLR scores, A further step of calibration is clearly needed. As discussed in the previous chapter, a logistic regression fusion might be a good option to effectively select pertinent attributes and obtain enhanced discrimination performance and well calibrated LLRs.

Finally, in this chapter we have exclusively examined Steps 1 and 2 of our approach. The application of Step 3 to the BAE-vectors is left for future investigation, offering the opportunity to uncover the encoded information within these attributes. Nevertheless, it is important to note that extra explainability analyses of the application of BA-LR on BAE-vectors as well as the application of Step 3 on SPINE-vectors is provided in appendix D.

Chapter 11

Conclusion and perspectives

Automatic speaker recognition (ASpR) systems have found their way into numerous applications, including security systems, access control, forensic investigations, and personalized assistant services. These systems utilize complex black box deep neural networks (DNN) and convey their outcomes through a single value. Despite their high performance, current ASpR systems fall short of providing an acceptable and satisfactory level of interpretability and explainability of the encoded speech representations and their role in the decision-making process. This opacity poses notable challenges in addressing ethical and legal issues, especially in critical domains like forensics, where the risk of introducing discrimination bias due to a black box DNN model, is a paramount concern.

This thesis introduced a three-step methodology based on deep learning, designed to achieve a trade-off between performance and providing interpretable and explainable ASpR results. This work has been positioned within a forensic context due to the critical requirement for interpretability and explainability in such settings. The goal of this work is to provide all stakeholders in the judicial process, including forensic practitioners and the court, with an interpretable and explainable assessment of the value of evidence. The concept behind the developed solution in this thesis was inspired by forensic DNA identification, renowned for its straightforward and easily understandable framework for identifying criminals. Drawing upon this inspiration and with all cautions considered, we engaged into a careful yet innovative analogy to introduce interpretability aspects into the ASpR process. This inspiration served as the dreamlike solution we aspired to achieve in this thesis. Our proposed methodology is composed of three steps, where each step is dedicated to add a further level of interpretability or/and explainability for ASpR system.

The **first step** aimed to extract and represent speech samples through attribute-based representations, modeled by a set of discriminant and independent voice attributes shared among groups of speakers. This representation is more easily understandable, allowing for a clear restructuring of the speaker information encoded within speaker embeddings. For this purpose, as a first attempt, we modified a SOTA ASpR model, specifically ResNet, to concentrate speaker information into distinct dimensions,

known as attributes, during the training for speaker classification. The experimental results demonstrated a slight decrease in ASpR performance compared to a baseline x-vector with approximately 2% in absolute EER on VoxCeleb1 dataset. Nevertheless, the obtained representations showed a better handle of the information encoded into a reduced number of dimensions than x-vectors, thereby achieving an acceptable trade-off between performance and interpretability.

Building upon these representations, the **second step** addresses the lack of interpretability and explainability in the process of score calculation for an ASpR task. In forensics, this score is represented by the Likelihood Ratio (LR), seen as the gold standard for evaluating evidence in legal proceedings. The goal is to provide a transparent process using DNN-based representations to estimate the LR. This enhances the confidence of the court in the value of evidence and allow to gain further insights into the factors influencing this value. To this end, we introduced BA-LR, a fully transparent and understandable framework for computing the LR value. This framework decomposes the LR estimation process into independent sub-processes, each dedicated to a particular attribute. The sub-processes are mainly attribute-LRs that are computed based on two forensic hypotheses: prosecution and defense. These attribute-LRs are explicitly estimated using attribute behavioral parameters such as typicality and uncertainty. In this work, we presented two versions of attribute-LR calculation. The first version is inspired by DNA, while the second is grounded in more reasonable and speech-based assumptions. It accounts for the likelihood of error on both sides of a comparison pair. The final log LR value is computed as the sum of attribute-log LRs, assuming independence between attributes. Evaluated on three different test corpora, namely VoxCeleb1, VOiCES and SITW, our solution demonstrated its generalisation abilities. In terms of ASpR results, BA-LR showed comparable performance to a baseline x-vector on all datasets for both versions, with an absolute average loss of 1.72% compared to the baseline, even though it uses a speech representation that is $\sim$40 times more compact. In terms of explainability results, BA-LR demonstrated inherent explainability by presenting attribute-log LRs akin to Shapley-like explanations. The obtained explanations revealed that the final LR is not arbitrary; instead, it is predominantly influenced by discriminant and rare attributes.

The **third step** focuses on explaining the encoded information in attribute-based representations. Specifically, we provided an automatic description of the nature of attributes encoded in the binary vectors, aligning them with voice characteristics. This description offers insights about the vocal information encoded by the DNN model and involved in the process of ASpR scoring. To achieve this, we introduced a novel methodology establishing a mapping between attributes and descriptors automatically extracted from speech, without requiring additional labeling or annotations. This fully automatic explainability method operates at both the utterance and frame levels. At the utterance level, we directly mapped extracted acoustic parameters from speech samples of the train set, VoxCeleb2, to binary vectors. The obtained phonetic descriptions of attributes using two different mapping methods, statistic SLDA and an inherently interpretable model, demonstrated convergence of up to 80%, enhancing trust in these descriptions. Also, evaluated on a test set, VoxCeleb1, these descriptions have shown their fidelity, generalisation and consistency. For frame-level representations, another

approach and additional processing was employed. Through backpropagation in the ResNet architecture, we demonstrated the ability to align each present attribute with its corresponding input frames. Thanks to this alignment, we were able to localize particular temporal information related to each attribute and provide a finer grained description in terms of phonemes. The phonemic descriptions showed a higher frequency of selection for vowels and nasals among attributes, with a more pronounced presence in certain attributes compared to others. The resulting descriptions and explanations revealed that attributes encode generally distinct phonetic and phonemic information. They also provided new combinations of phonetic and acoustic features, serving as an informative tool for phoneticians.

Building upon these three steps, we presented an application of BA-LR framework in a forensic context using the NFI forensically realistic database, NFI-FRIDA. The aim is to further evaluate our approach and to handle the existing mismatch between training conditions of VoxCeleb2 dataset and the evaluation forensic scenarios. In this context, we developed a global Logistic Regression model that effectively calibrate the final LLRs. A fusion approach of attribute-LLRs using sparse logistic regression was also introduced to select only significant set of attributes involved in the final LLR computation. The overall ASpR performance obtained on NFI-FRIDA proved the generalization power of BA-LR, even though the BA-extractor model and BA-related parameters were trained on a different language and condition far from the forensic ones. Compared to baseline x-vector, an average slight increase in absolute EER of 0.85% for all devices is observed using BA-LR, except for device 4 of forensic conditions with 1.66% increase of average absolute EER. The Logistic-Regression based calibration approach showed its abilities to produce well calibrated LLRs, even when the mismatch with the training set was particularly large. The fusion approach enabled us to regulate any potential correlation between attributes. It offered significant performance gains in difficult scenarios, occasionally surpassing x-vectors. This was achieved thanks to its ability to completely eliminate the influence of certain BAs, particularly affected by domain mismatches. As expected, this Logistic Regression based fusion also provided a level of calibration equivalent to the global calibration.

Finally, we added one improvement over the first step of our approach, aiming to address some limitations of the initial BA-extractor. Firstly, the objective of shared attribute modeling is not explicitly taken into account into the BA-extractor. Secondly, binarization is not directly involved into the modeling, but applied after the training of the model. To address these limitations, we proposed two solutions based on auto-encoder model. The first is SPINE auto-encoder that encourages sparsity in the generated vectors. Experimental analyses revealed that despite its comparable ASpR performance to x-vectors, a 0.29% increase of absolute EER of SPINE on VoxCeleb1 using cosine similarity, this model fails to accurately model attribute behavior in the representations. Hence, it is shown to be incompatible with BA-LR framework. As a second solution, we proposed a binary auto-encoder, BAE model, that includes a dedicated loss function pushing to model attribute-based behavior in the binary representations. Experiments on VoxCeleb1 showed the effectiveness of our BAE model, with an absolute average reduction in EER of 1.39% compared to the BA-extractor, from 3.7% to 1.8% of EER, while offering the same level of explainability. These results

are not only promising but they also underscore the high potential of our approach to achieve an excellent trade-off between performance and explainability/interpretability.

Overall, the three-step method[1] introduced in this thesis opens a new perspective on explainable and interpretable ASpR systems. It provides a practical tool for better understanding the intricate information encoded by DNN models. While its applicability extends beyond forensic scenarios, its importance is particularly notable in forensics, where the interpretability of the results often outweighs performance. In this regard, our approach serves as a valuable resource for forensic practitioners, shedding light on the inner workings of DNN-based ASpR systems and the factors influencing their outputs and helping them to discover more about vocal information. It also holds great potential for aiding the court in making well-informed decisions.

Perspectives and future work

The results obtained in this thesis present a promising direction for future work in interpretability and explainability for ASpR systems. Given that future work is discussed at the end of each chapter, we suggest here more global perspectives related to the entire work. These perspectives include suggestions for enhancing performance and extending the application of our three-step approach to other domains and contexts.

Taking a broader perspective on recent advancements in DNN speaker models and their performance, we believe that the binary-attribute-based extractor of Step 1 has the potential to achieve performance comparable to SOTA models. One avenue of improvement is to use the recent speaker architectures such as ECAPA-TDNN or MFA-conformer and incorporate both, the binary aspect and the attribute-oriented aspect into the generation of speaker embeddings. This could be done either in a speaker classification task, or in an auto-encoder fashion by training from scratch or by fine-tuning. In the former solution, the STE technique, the batch organization and the attribute-oriented loss are added. In this case the model is trained to classify speakers using the generated binary vectors. For the latter, it is the architecture of the BAE encoder that could be replaced by one of the SOTA architecture, while maintaining the same auto-encoder. Another perspective on this extractor to be investigated, is inspired from Vector quantization. Instead of directly generating binary vectors, this extractor might be an auto-encoder that is designed to learn a set of separated clusters of dimensions in the latent space. Each cluster of dimensions form an attribute. Then a binary vector is composed having as dimensionality the number of clusters where each dimension indicates if an attribute is present in the corresponding cluster or not.

Even though the dimensions of all the proposed extractors in this thesis have been experimentally demonstrated to be well decorrelated, the independence assumption is not explicitly enforced as a constraint during the training of any of them. This aspect remains an area for future exploration and improvement. As a potential solution, we suggest that the incorporation of decorrelation or independence objective functions during training such as Barlow Twins [312] or Hilbert-Schmidt Independence Criterion

[1]The code for most of our work is available on GitHub: https://github.com/LIAvignon/BA-LR

[2] would be a good option.

Similar to the work conducted in the JSALT2023 workshop context in Appendix D, we believe that our approach has the potential to offer insights into the encoding of various voice characteristics. For instance, when investigating the characteristics contributing to nasality, voice quality, or any other attribute, the third step of our approach can be utilized to uncover the relevant acoustic and phonetic parameters encoding these characteristics. Our approach stands as a valuable tool for phoneticians, providing a means to understand certain aspects of vocal information. Another ongoing PhD work is currently exploring this aspect further, aiming to uncover combinations of phonetic features as high-level features using our approach.

An interesting avenue for investigation involves the application of our work to a **text-dependent** speaker recognition task, where the recognition requires that the speaker utter specific predefined words or passphrases. Such task is commonly employed for authentication purposes in applications like banking. Given that banking applications fall into the category of high-stakes contexts, this explainable and interpretable application would be beneficial. Implementing our three-step approach in this context, where linguistic content is predetermined, allows not only to better map attributes with phonemes but also to focus more on other speaker variabilities while eliminating content-related variability. For instance, this would inform us whether the recognition of a speaker is based on his pronunciation of specific phonemes.

Another pertinent application of our approach lies in the domain of **speaker diarization**, tasked with determining "who speaks when?". By modifying the hypotheses employed for likelihood ratio calculation, where the prosecution assumes the speech sample belongs to speaker A and the defense hypothesis posits it belongs to speaker B, likelihood ratios can be computed for each pair of speakers in a speech segment. Using the interpretable BA-LR framework for these likelihood ratios allows to explore the reasons behind transitions to different speakers at specific points within the speech segment.

The obtained descriptions of attributes in this work can also serve **privacy-related** tasks by enabling the hiding of personal information about the speaker. Understanding the nature of attributes facilitates the identification and suppression of specific attributes directly linked to sensitive characteristics, thereby enhancing privacy protection.

Finally, we also believe that our approach is not restricted to speech and could be applied on other types of data. For instance, exploring its application in **forensic text comparison** could unveil specific characters or distinctive writing styles that differentiate individuals. If adhering to the predefined criteria established for the extraction of binary vectors to represent samples, we believe that the same process described in this work could be applied to text, images, or any other data type.

[2] https://jejjohnson.github.io/research_journal/appendix/similarity/hsic/

Part IV

Appendices

Appendix A

Extracts from Case text and judicial articles

A.1 Courts positions in use cases

A.1.1 Over reliance position

Superior Court of Pennsylvania

A summary from People v. Lang: In Commonwealth v. Foley (Pa. Super. Ct. 2012) 38 A.3d 882, 890 (Foley), a Pennsylvania appellate court upheld admission of testimony based on the interpretation of data using the TrueAllele program against a challenge that it did not meet the Frye standard[1].

Some extracts from the case text are the following: *[...] Pennsylvania continues to adhere to the Frye test, which provides that "novel scientific evidence is admissible if the methodology that underlies the evidence has general acceptance in the relevant scientific community." [...]The trial court did not expressly determine whether Dr. Perlin's testimony was "novel scientific evidence." Opinion and Order of Court, March 3, 2009, at 2–3. Instead, the court found that Dr. Perlin's methodology was a refined application of the "product rule," a method for calculating probabilities that is used in forensic DNA analysis. See id., at 2. The Pennsylvania Supreme Court has held that scientific evidence based on the product rule is admissible in the Commonwealth. See Commonwealth v. Blasioli, 552 Pa. 149, 713 A.2d 1117, 1118 (1998). Because Dr. Perlin's calculations were made using newer technology, the trial court rhetorically asked "at what point does the use of the product rule become novel science." Opinion and Order of Court, March 3, 2009, at 2. The trial court went on to find that Dr. Perlin's methodology was generally accepted. [...] we find no legitimate dispute regarding*

[1]The Frye standard is a legal precedent that determines the admissibility of scientific evidence in court.

the reliability of Dr. Perlin's testimony. Dr. Perlin used proprietary software called TrueAllele to interpret the data he received from the FBI. See N.T., March 12, 2009, at 130.

UNITED STATES COURT OF APPEALS FOR THE SIXTH CIRCUIT

A summary from People v. Davis: Noting that there were more than 50 published peer-reviewed articles addressing STRmix at the time of the evidentiary hearing, and that one expert opined that it was the most tested and peer-reviewed probabilistic genotyping software available

A.1.2 No trust position

Supreme Court, New York County

[...] In each case the defense challenged the introduction of DNA evidence created by the Forensic Statistical Tool ("FST"). The FST was an analytic tool with which the city's Office of the Chief Medical Examiner ("OCME") assigned "likelihood ratios" to forensic samples made up of DNA from not one, but two or three, individuals. A scientist could use FST results to opine that a two-person DNA mixture was X times more (or less) likely to be made up of DNA from a particular known individual and one unknown, unrelated individual than DNA from two unknown, unrelated individuals. Similarly, the analyst could testify that a three-person mixture was X times more (or less) likely to be from a particular known individual and two unknown, unrelated individuals than from three unknown, unrelated individuals.

[...]The defendants' challenges asserted that the FST results were not the product of procedures generally accepted in the "community" of DNA forensic scientists. This court once before faced such a claim, and ruled after a Frye hearing that FST results should indeed be excluded on that ground. People v. Collins , 49 Misc 3d 595 (Sup Ct Kings Co 2015). For this case, this court assessed whether developments in the community of forensic scientists since the Collins decision of July 2, 2015, should change that conclusion. For the reasons noted below, this court decided on October 16, 2017 that it should again exclude the challenged FST evidence. The court has continued reviewing developments as best it can in the period since, seeing no basis to re-think the conclusion that nothing has changed.

A.1.3 Reasonable reliance position

The Supreme Court of the United States

A summary of the case from "BRIEF FOR THE AI NOW INSTITUTE, AMERICAN CIVIL LIBERTIES UNION, ELECTRONIC FRONTIER FOUNDATION, CENTER

ON RACE, INEQUALITY, AND THE LAW, AND KNIGHT FIRST AMENDMENT INSTITUTE AS AMICI CURIAE SUPPORTING THE RESPONDENT": Governments are also implementing automated decision making systems to evaluate the performance of employees including, for example, public school teachers. A school district in Texas implemented one such "data driven" teacher evaluation model through privately developed software that purported to compare the results of a teacher's students to classroom statistics across the state and within the teacher's prior performance record. Teachers sued the district, arguing that the software was fundamentally inscrutable and that there was no way for teachers to know whether the software was accurately assessing their job performance. The court agreed, holding that the "teachers have no meaningful way to ensure correct calculation of their [evaluation] scores, and as a result are unfairly subject to mistaken deprivation of constitutionally protected property interests in their jobs." Id. at 1180. Similar systems purporting to measure the efficacy of government employees are likely to proliferate. Without meaningful transparency, these systems will raise serious concerns about fairness and accuracy.

A.2 Judicial articles

A.2.1 Article 6 from the European Court of Human Rights

Right to a fair trial

In the determination of his civil rights and obligations or of any criminal charge against him, everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal established by law. Judgment shall be pronounced publicly but the press and public may be excluded from all or part of the trial in the interests of morals, public order or national security in a democratic society, where the interests of juveniles or the protection of the private life of the parties so require, or to the extent strictly necessary in the opinion of the court in special circumstances where publicity would prejudice the interests of justice. Everyone charged with a criminal offence has the following minimum rights:

- *a) to be informed promptly, in a language which he understands and in detail, of the nature and cause of the accusation against him[...];*
- *e) to have the free assistance of an interpreter if he cannot understand or speak the language used in court.*

A.2.2 Article 149 from The Belgian Constitution

Each judgment is supported by reasons. It is made public according to the terms specified by the law. In criminal matters, the operative part is pronounced publicly.

A.3 GDPR articles

A.3.1 Article 15: Right of access by the data subject

The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:

- *(a) the purposes of the processing;*
- *(b) the categories of personal data concerned;*
- *(c) the recipients or categories of recipients to whom the personal data have been or will be disclosed, in particular recipients in third countries or international organisations;*
- *(d) where possible, the envisaged period for which the personal data will be stored, or, if not possible, the criteria used to determine that period;*
- *(e) the existence of the right to request from the controller rectification or erasure of personal data or restriction of processing of personal data concerning the data subject or to object to such processing;*
- *(f) the right to lodge a complaint with a supervisory authority;*
- *(g) where the personal data are not collected from the data subject, any available information as to their source;*
- *(h) the existence of automated decision-making, including profiling, [...], at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.*

A.3.2 Article 22: Automated individual decision-making

- *1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*
- *2. Paragraph 1 shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent.*

- *3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.*
- *4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data [...] applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.*

A.4 AI act: Recital 38

Furthermore, the exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial as well as the right of defence and the presumption of innocence, could be hampered, in particular, where such AI systems are not sufficiently transparent, explainable and documented. It is therefore appropriate to classify as high-risk a number of AI systems intended to be used in the law enforcement context where accuracy, reliability and transparency is particularly important to avoid adverse impacts, retain public trust and ensure accountability and effective redress.

A.5 The Equal Credit Opportunity Act

It shall be unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction.

- *(1) on the basis of race, color, religion, national origin, sex or marital status, or age (provided the applicant has the capacity to contract).*
- *(2) because all or part of the applicant's income derives from any public assistance program.*
- *(3) because the applicant has in good faith exercised any right under this chapter.*

Appendix B

Appendix to Step 2

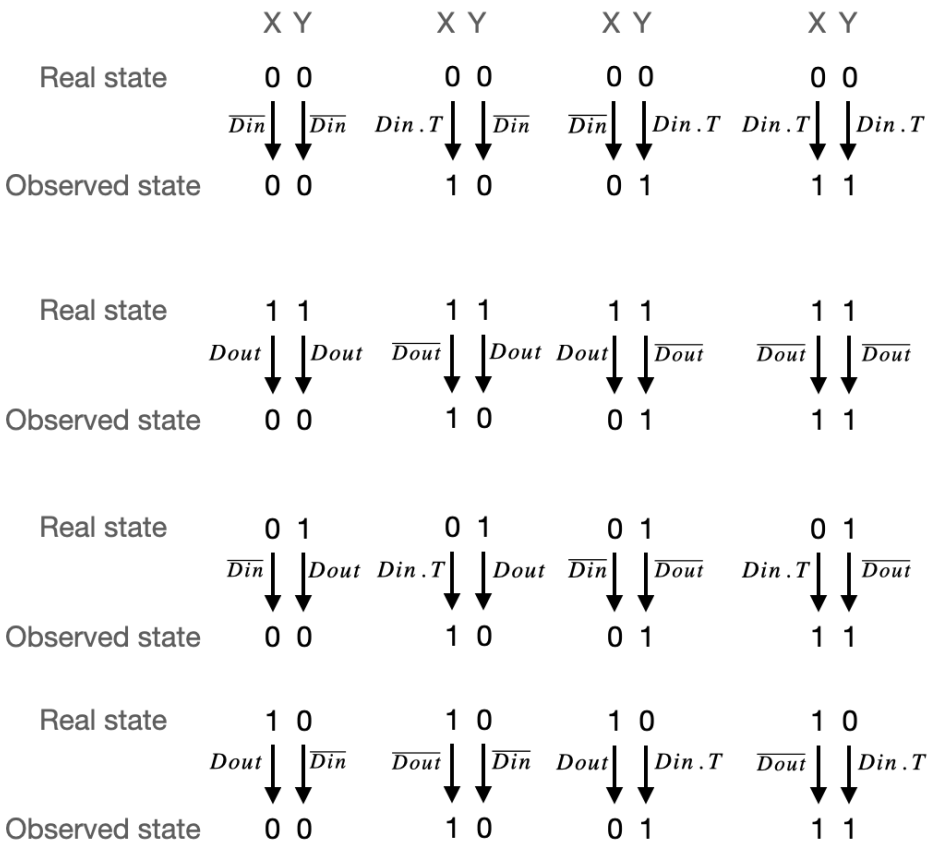


Figure B.1: All possible combinations of observed and real (i.e. actual) state for Speech-based version of BA-LR

Appendix C

Extra analyses and results of Step 3

C.1 Details of the BA models

Table C.1: Further details about the number of speech extracts selected for the first 16 BAs (out of 205) for train and test datasets.

	Train[1]	Test[1]	Tree depth	Accuracy train	Accuracy test
BA$_2$	141,998	45,759	8	0.70	0.60
BA$_3$	28,090	63,068	6	0.73	0.59
BA$_4$	89,254	51,830	10	0.77	0.64
BA$_5$	278,619	36,604	9	0.90	0.73
BA$_8$	160,936	23,923	9	0.67	0.60
BA$_9$	210,359	44,027	11	0.95	0.90
BA$_{10}$	39,603	39,360	9	0.67	0.56
BA$_{11}$	64,687	43,540	9	0.62	0.53
BA$_{12}$	140947	37,157	9	0.75	0.62
BA$_{13}$	131,764	23,534	10	0.68	0.62
BA$_{15}$	158,655	31,442	9	0.74	0.57
BA$_{16}$	37,541	62,536	8	0.61	0.61
BA$_{17}$	58,132	24,373	9	0.73	0.61
BA$_{18}$	143,791	21,413	10	0.70	0.63
BA$_{19}$	155,179	39,262	9	0.75	0.76
BA$_{20}$	117,696	14,997	8	0.87	0.59
...	...	...	...	..	...

[1] 1 Number of speech extracts per set. Number of speech extracts is balanced for S_0 and S_1

C.2 Correlation between attributes in terms of frames

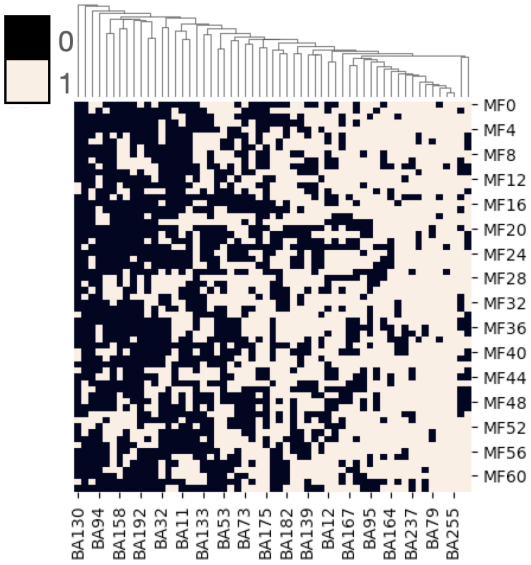


Figure C.1: A heatmap illustrating the binarization of the Softplus-matrix A for a given utterance, with clustering of BAs in terms of MegaFrames using Jaccard distance.

Figure C.1 presents a heatmap of a binarized version of the Softplus-matrix A for a given utterance. The binarization employs the same threshold used to derive BA-vectors from Softplus-vectors. A value of 1 signifies the selection of the MF by the BA, while a value of 0 indicates that the MF is not selected. This figure presents also a clustering of BAs in terms of MFs using Jaccard distance. This distance is calculated between two binary vectors as the ratio of the intersection divided by the union of elements. It is evident that the MFs are not uniformly selected by the BAs. Each BA demonstrates a unique representation in terms of frame-level units. However, distinctions can be observed among groups of BAs, with some selecting a larger number of MFs compared to others that select a more limited set of MFs.

Correlation between attributes in terms of MegaFrames

For a clearer understanding of the shared temporal information between attributes, we present in Figure C.2 the correlation between attributes across VoxCeleb1 utterances. The procedure is as follows: for each utterance, we select pairs of present BAs, calculate the intersection and union between them in terms of MFs, and then accumulate these intersections and unions across all utterances where the pair of attributes is present. The final correlation matrix between BAs is therefore calculated as the sum of intersections divided by the sum of unions of MFs as expressed in the following equation:

$$\text{Corr}_{BA_i,BA_j} = \frac{\sum \text{Intersection}(BA_i, BA_j)}{\sum \text{Union}(BA_i, BA_j)} \tag{C.1}$$

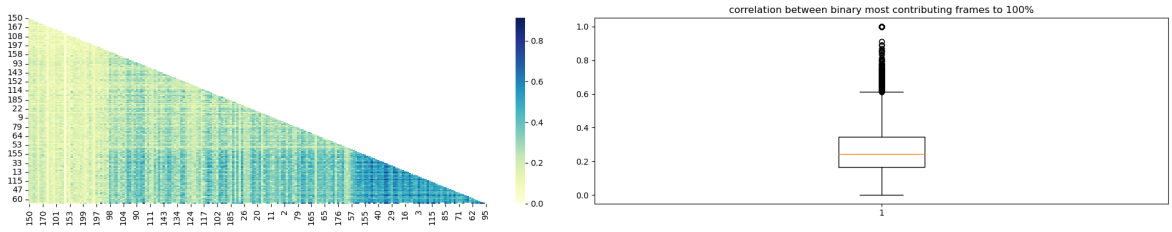


Figure C.2: Correlation between attributes in terms of MegaFrames contributing to 100%

C.3 Alignment of MegaFrames with phonemes for an attribute

Thus far, we have introduced a temporal characterization of attributes, highlighting that attributes exhibit distinct patterns in terms of MegaFrames. In this section, our objective is to delve into the information encoded in these MegaFrames to offer a temporal description of attributes. To achieve this, we establish an alignment between MegaFrames selected by a specific attribute and input frames by retracing the flow through the ResNet extractor. This alignment enables the identification of frames related to a particular attribute. Figures C.3, C.4, C.5, C.6,C.7 show the MF activations for five different BAs for the same portion of utterance. These figures show that for each BA, the activated MFs are not behaving similarly. For instance, $BA_1 1$ reflects the activation of MF2 and MF3 while $BA_1 7$ presents null activation for that part of utterance. This might be explained by the fact that this portion of the utterance does not contain any pattern detected by $BA_1 7$.

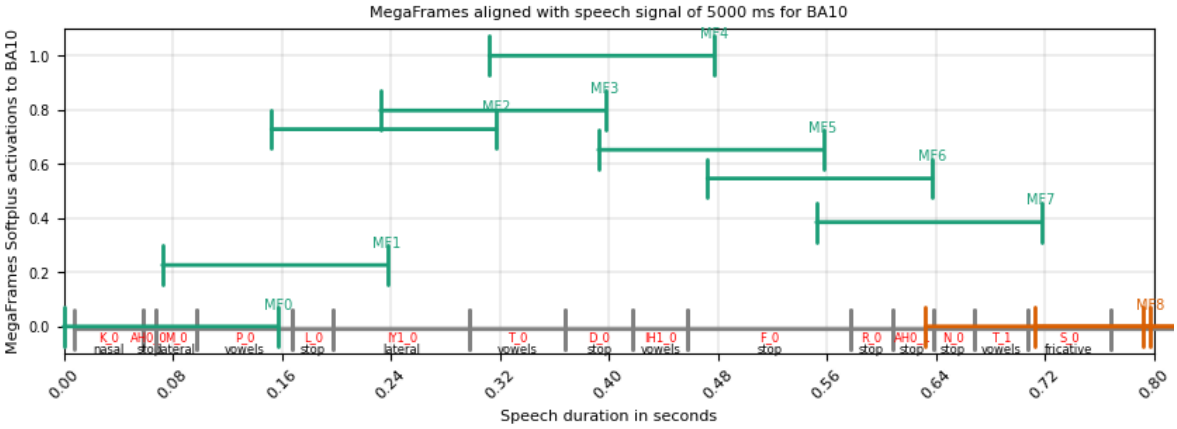


Figure C.3: MegaFrames activations to BA_{10}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes

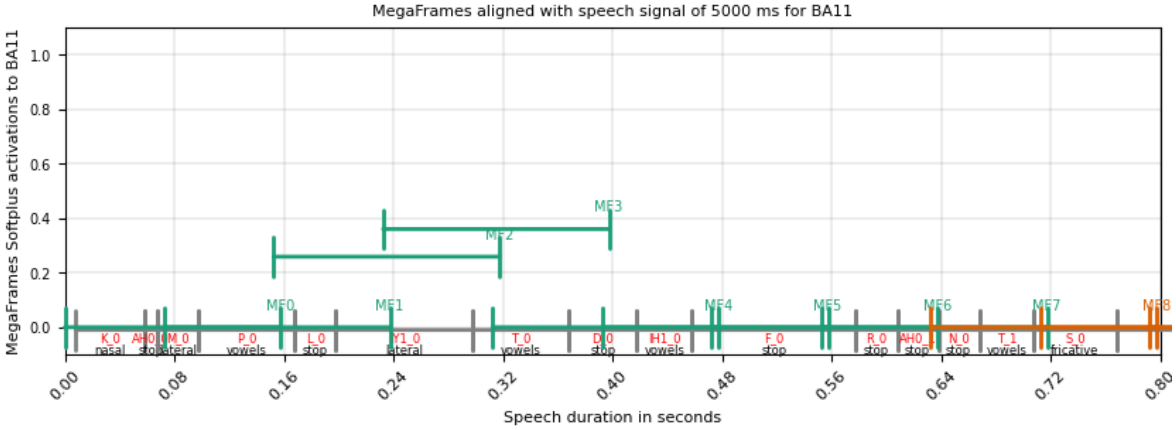


Figure C.4: MegaFrames activations to BA_{11}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes

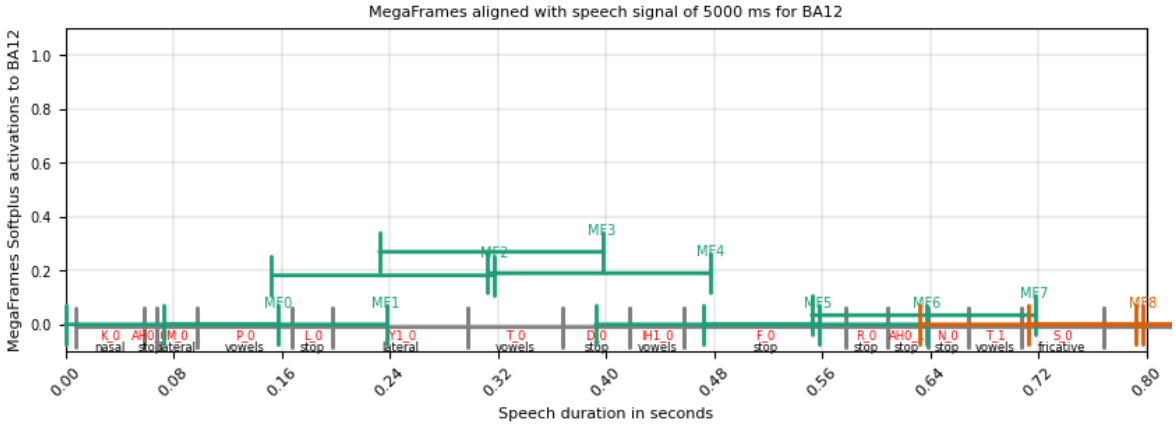


Figure C.5: MegaFrames activations to BA_{12}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes

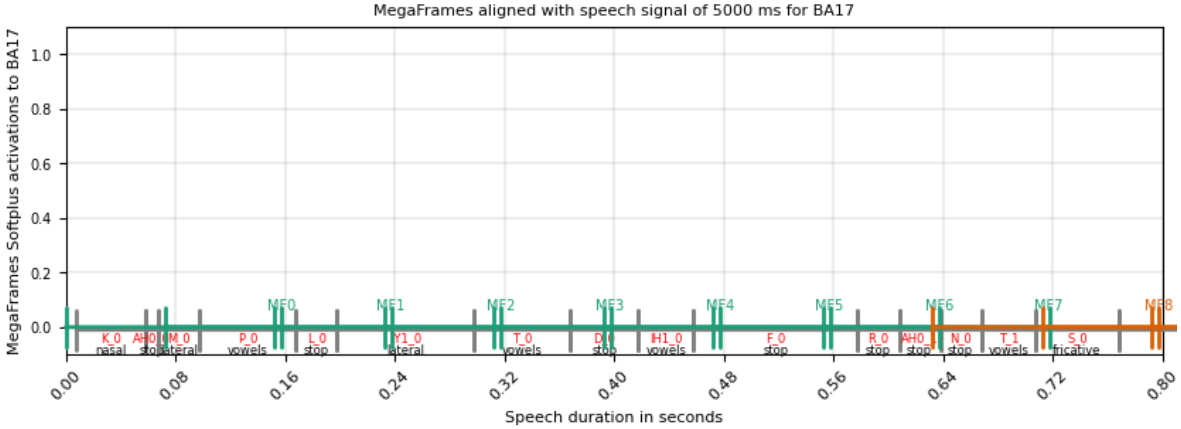


Figure C.6: MegaFrames activations to BA_{17}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes

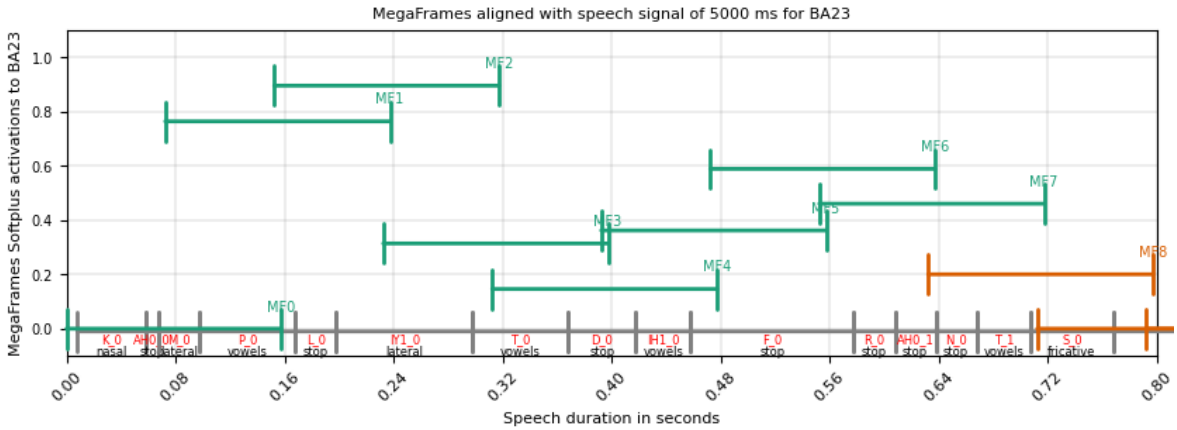


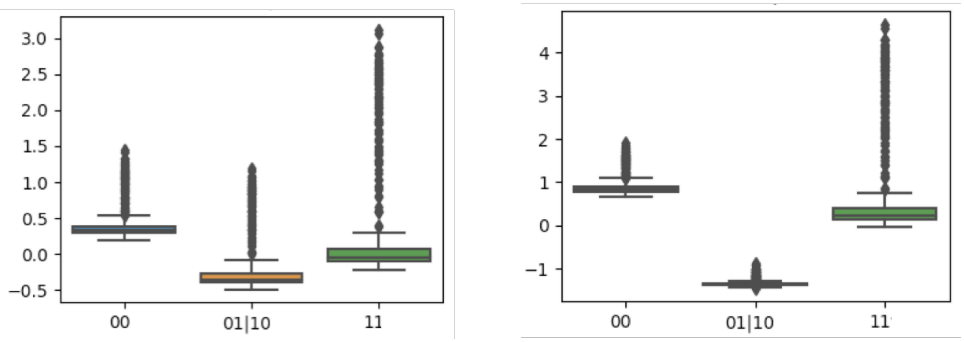
Figure C.7: MegaFrames activations to BA_{23}=1 in a portion of 0.8s of a speech utterance of 5s aligned with phonemes and classes of phonemes

Appendix D

Extra results and analyses of BAE and SPINE vectors

D.1 Extra explainability analyses of BA-LR scoring on BAE system

Thus far, we have demonstrated that applying the BA-LR framework to the proposed BAE model yields in superior speaker recognition performance compared to the BA-extractor. In this section, our objective is to delve into some explainability aspects pertaining to the calculation of LLRs using different versions of BA-LR. To achieve this, Figure D.1 illustrates the distributions of different attribute LLRs types applying both versions of BA-LR.

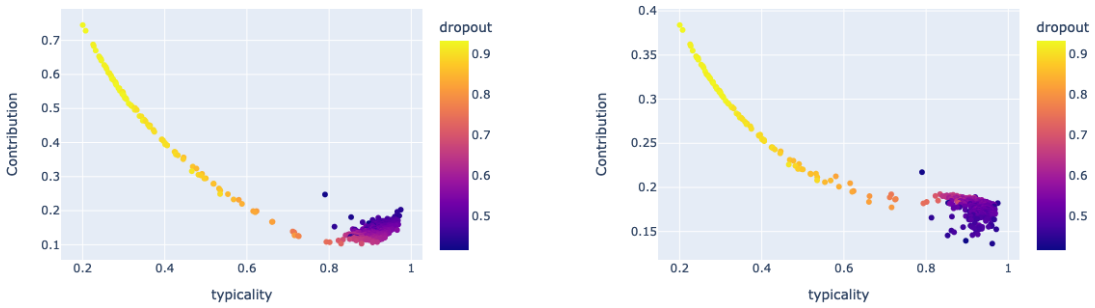


(a) Using DNA-inspired (b) Using Speech-based

Figure D.1: Distribution of attribute LLRs using BAE-vectors

Clearly, the attribute LLR values in Speech-based are more explainable and reasonably distributed compared to the attribute LLRs in DNA-inspired. For instance, the attribute LLRs 11 exhibit positive values in Speech-based, while in DNA-inspired, some values display slightly low negative values. Similarly, for attribute LLRs 01|10,

Speech-based exclusively presents negative values, whereas in DNA-inspired, certain attribute LLRs are observed to be positive which should not be the case.



(a) Using DNA-inspired

(b) Using Speech-based

Figure D.2: Relationship between behavioral parameters and the contribution of attribute LLRs using BAE-vectors

Figure D.2 depicts the contributions of BAE attributes in the two versions of LLRs computation, correlated with the behavioral parameters. It is crucial to note that these contributions represent the attribute LLR values themselves. As expected, attributes with low typicality [0.2, 0.4] exhibit the highest contributions to the final LLR, falling within the range [0.4, 0.7] and [0.25, 0.4] for DNA-inspired and Speech-based, respectively. The vast majority of attributes demonstrate very high typicality, ranging from 0.6 to 1, resulting in low contributions within the range [0.1, 0.2] for both DNA-inspired and Speech-based. Given that drop-out is more pronounced for lower typicality values, attributes with the highest contribution exhibit higher drop-out compared to others.

D.2 An additional investigation of Step 3 on SPINE-vectors in JSALT workshop

This section focuses on explaining the SPINE speaker embeddings using the Step 3 methodology of our approach. This work was done in JSALT workshop2023. In the next sections, We evaluate the performance of these representations in a speaker recognition task. Subsequently, we delve into the explainability schema proposed to explore information in these binary representations, adopting our Step 3 methodology. Finally, we conclude with a discussion and potential future directions.

D.2.1 Evaluation of SPINE representations

Figure D.3 presents the speaker recognition performance of the three SPINE systems on VoxCeleb1 trials for a speaker recognition task, in terms of EER. In contrast to the less

sparse system, SPINE-15%, which exhibits a minimal discrimination performance loss of approximately 0.29% compared to the baseline x-vector (1.37%), the loss becomes more pronounced with higher sparsity levels. Notably, the binarized version of the SPINE-vectors shows a negligible loss, maintaining nearly the same ASpR performance as the sparse vectors.

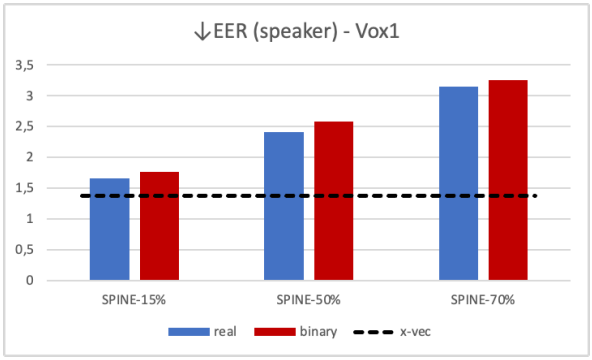


Figure D.3: EER% using SPINE-vectors on VoxCeleb1

D.2.2 Explainability of SPINE representations

In this section, our goal is to explore the information encoded within specific dimensions of SPINE binary vectors, particularly those that are significantly important for two probing tasks. These dimensions are later referred to as features.

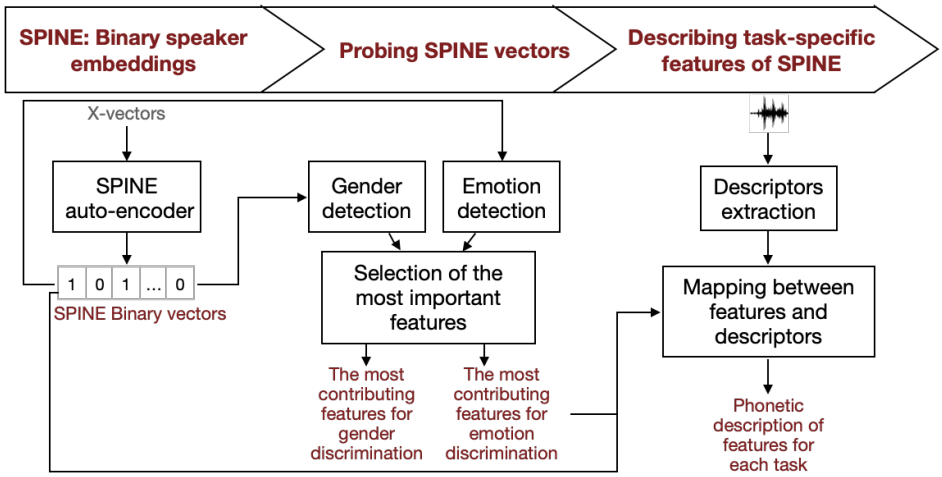


Figure D.4: An overview of the explainability schema for SPINE representations

Figure.D.4 provides an overview of the explainability framework for SPINE representations. Using SPINE binary vectors, we firstly assess their performance in two probing tasks: gender and emotion detection. Subsequently, we conduct a feature selection process to identify the most influential features for each classification task. Employing our three-world explainability method, these *task-specific features* are highlighted within the SPINE binary vectors, establishing a mapping between these features

and certain phonetic descriptors. Consequently, this facilitates the extraction of a phonetic description for these features. The subsequent sections delve deeper into the exploration of each block in this process.

Probing and features selections of SPINE representations

In this section, SPINE-vectors and their binary version are probed for gender and emotion detection, in order to investigate the presence of this information in these vectors. Following that some selection methods are used to select the most contributing features of SPINE-vectors to each task.

Emotion and gender classifiers are trained using a RandomForest (RF) model with cross validation. Two datasets are used to evaluate each task, namely VoxCeleb1 for gender and IEMOCAP [313] for emotion, where 80% of data is dedicated for train and 20% for test. IEMOCAP[1] is an acted, multimodal and multispeaker English database and it is annotated into 4 classes such as anger, happiness+excitement, sadness, neutrality. The number of samples per class for each task is illustrated in Table.D.1.

Table D.1: Number of samples per class for emotion and gender detection

	Emotion classes				**Gender classes**	
	Anger	Happiness	Sadness	Neutral	Male	Female
#Samples	1103	1636	1084	1708	90450	63066

Table.D.2 presents the performance of SPINE-vectors, their binary version and x-vectors for gender and emotion detection tasks.

Table D.2: Evaluation of performance in emotion and gender detection tasks

	Emotion (UAR)	**Gender (AUC)**
Datasets	**IEMOCAP**	**VoxCeleb1**
Binary	49%	97%
Sparse	54%	99%
X-vector	55%	98.5%

For gender detection, accuracy is employed as the evaluation metric, while emotion detection utilizes the unweighted average recall (UAR). The UAR calculation involves considering the recall (sensitivity) for each emotion class independently and then averaging these values without weighting. Emotion detection proves to be a challenging task for x-vectors, yielding performance close to randomness, when using SPINE sparse and binary vectors. Significantly, SPINE-vectors are shown to be good gender detectors, with sparse vectors exhibiting a slight improvement over x-vectors. This performance

[1]https://sail.usc.edu/iemocap/

distinction underscores the strong presence of gender information in SPINE-vectors, while indicating a more subtle presence of emotion information.

To select the most contributing features of SPINE-vectors to each classification task, we choose three different methods: Gini Index of RF [314], Shap [199] and test statistic SLDA [264, 265]. The use of three distinct methods aims to enhance confidence in the selected features. In the following, we review the used selection methods:

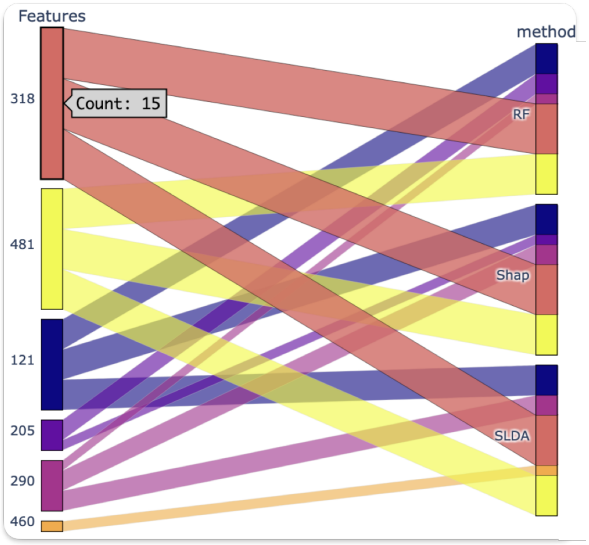


Figure D.5: Most important features for gender detection selected using three methods

- *Shap*: Shapley values are calculated considering one feature at once along with the impact of all its permutations on the RF model output.
- *RF (Gini index)*:The Gini Index is a measure of impurity or disorder within a node of a decision tree, and it is employed in Random Forests to determine the best split for a node based on the feature that maximally reduces impurity. It does not inherently account for feature correlations.
- *SLDA*: It is not based on the RF model, but it quantifies the discriminant power of each feature independently of others with respect to the output class.

Figure D.5 depicts the six most crucial features for gender detection task, as chosen by the three selection methods. The size of the lines matching features with methods indicates the rank attributed to the feature by the corresponding method, with larger lines indicating a higher importance ranking. The large lines in the figure signify that the first three features are identified as the most important by all three methods. Conversely, the last three features are shown to be less significant and exhibit lower confidence, as not all methods converge in their ranking.

The task of selecting the most influential features for emotion detection proved to be highly challenging, with no convergence observed among the three methods. This was expected, given the difficulty of accurate emotion detection and the limited

discriminant information between emotion classes. As a result, we opt to proceed with the features selected by the SLDA method (i.e F290, F329, F192, F342), for the subsequent exploration.

Mapping between SPINE features and descriptors

In this section, our objective is to explore the phonetic information encoded in these selected task-specific features. To accomplish this, we employ the three-world explainability method introduced in Step 3 of our approach. This involves establishing a mapping between these features, which represent the D world, and the same set of descriptors used in Step 3 (eGeMAPs)[266], forming the I world. It is essential to note that this mapping is performed individually for each feature, employing phonetic descriptors to discriminate between the two classes (0 or 1) of the feature, as illustrated in Step 3 methodology.

To enhance confidence in the phonetic description of task-specific features, various mapping functions are employed to select most relevant descriptors for each feature, including a surrogate model such as RF followed by Shap or by Gini index, the SLDA, and a Mutual Information-based method known as Double Input Symmetrical Relevance (Disr) [315]. The first three methods have been previously described and utilized in the earlier task, while the latter is introduced here for the first time. Disr considers that the combination of descriptors provides more information about the output class than considering each descriptor individually.

Table D.3: Families of descriptors

Family	**Descriptors**
F0	F0, logRelF0
MFCC	MFCCs 1, 2, 3, 4
F1	F1 parameters
F2	F2 parameters
F3	F3 parameters
Spectral	alpharatio, slope, spectralFlux and hammarbergIndex
Rythm	Unvoice/voiced parameters
Loudness	loudness parameters and equivalent-sound level
VoQ	jitter, shimmer and HNR parameters

Figure.D.6 and Figure.D.7 show the phonetic description of task-specific features using the selected descriptors grouped by families. The grouping of descriptors into families is proposed in Table.D.3. The size of the lines liking features with descriptors and families reflects the level of convergence in ranking among the four methods, where larger lines indicate a higher degree of consensus in their selections. Figure D.6 reveals that gender-specific features are predominantly described by two significant families: F0 and MFCC. For instance, Feature 318 encodes information primarily related to

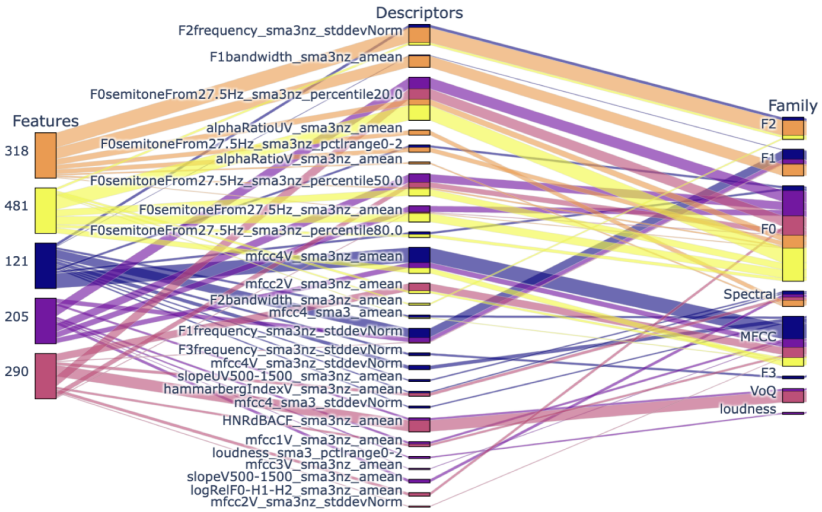


Figure D.6: Phonetic description of gender-specific features, grouped by families of descriptors

formants F1 and F2 families, along with the F0 family. Feature 481 is rich in information related to F0, while Feature 121 encodes more information about MFCC and F1 families. Feature 205 is predominantly associated with F0, while Feature 290 is more closely related to voice quality (VoQ) and F0. This description highlights the importance of F0 and MFCCs for the task of gender detection.

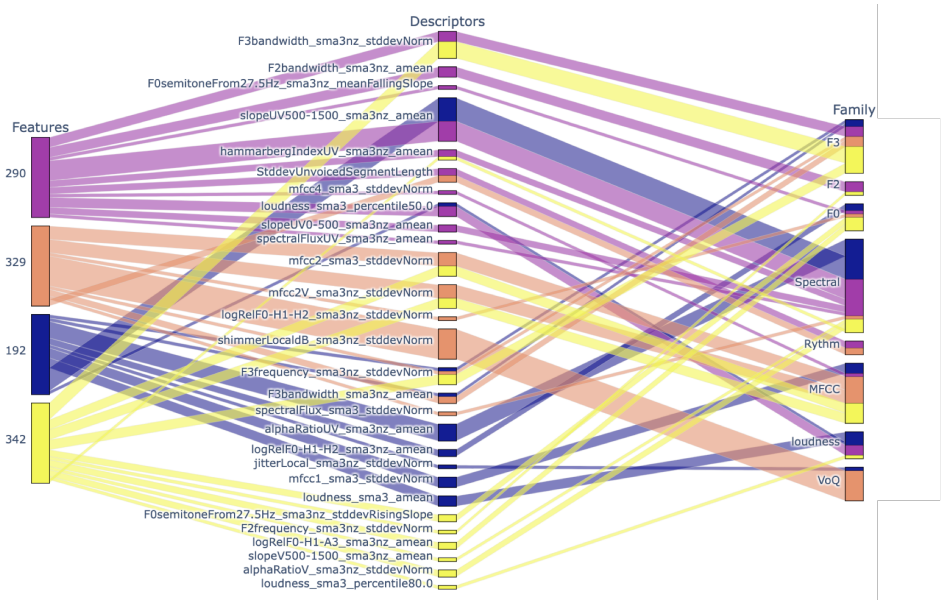


Figure D.7: Phonetic description of emotion-specific features, grouped by families of descriptors

Figure D.7 illustrates emotion-specific features, primarily characterized by F3, spectral parameters, MFCC, loudness, rythm and voice quality. For example, Feature 290 is primarily encoding spectral characteristics with high confidence, along with F3, F2,

rhythm, and loudness. Feature 329 confidently encodes voice quality, along with some MFCCs and rhythm, while Feature 192 predominantly encodes spectral characteristics with high confidence, in addition to MFCC and loudness. Feature 342 is more distinctly described by the F3, F0, spectral features, and MFCCs familes. The description of these features demonstrates the importance of prosody and spectral characteristics for emotion detection.

The disparity in descriptors selection by the four methods for certain features could be attributed to the high correlation among eGeMAPS descriptors [266] (Figure.D.8). This correlation introduces divergence in the chosen descriptors by each method, as each method employs its unique approach to handle selection.

D.2.3 Discussion & perspectives

We introduced an alternative perspective of the Step 3 methodology from our approach, specifically the three-world method, aiming to enhance the explainability of speaker embeddings. The main goal is to explore SPINE-vectors and describe the encoded information. This is accomplished by emphasizing certain task-specific features within the dimensions of the embeddings and providing a phonetic mapping with these features. Our three-world explainability method proves highly beneficial in establishing this mapping and delivering a meaningful description of these task-specific features.

Advantages of SPINE binary vectors and their explainability

SPINE-vectors holds some significant advantages, summarized as follows:

- *General aspect*: It is demonstrated in this chapter that Step 3 of our approach is general and totally independent of Step 1 and 2 and that it could be applied to provide explanations on any other binary or discrete representations.
- *Binarization does not impact encoded information* SPINE-vectors exhibit good performance in gender detection and achieve results close to x-vectors in emotion detection task. The binarization aspect of SPINE-vectors is definitely useful and facilitates the application of the three-world explainability method.
- *Reliable feature selection*: The use of different selection methods reinforces the confidence in the selected specific features for each task.

Criticisms and perspectives of SPINE binary vectors

However, from a critical perspective, we acknowledge that this work does not encompass certain crucial aspects, leaving room for various future perspectives to enhance it, as highlighted below:

- *Trade-off sparsity Vs. performance*: In terms of performance, SPINE-vectors do not surpass BA-vectors. Specifically, under the same sparsity level, SPINE-70% vectors exhibit performance closely comparable to BA-vectors.

- *SPINE dimensions are not attributes !*: While SPINE-15% vectors demonstrate notable performance, it is essential to note that their success does not necessarily imply that these binary vectors can be unequivocally regarded as binary-attribute-based representations. In contrast, SPINE dimensions lack an attribute-like behavior. This renders the application of the BA-LR framework on SPINE-vectors impractical. This observation emphasizes the considerable challenge in extracting binary-attribute-based speaker embeddings, justifying the criteria established in Step 1 of our approach.

- *Inaccurate probing classifier, inaccurate features !*: It is crucial to emphasize that a probing task exhibiting poor performance suggests that the discriminative information between classes is weak and not explicitly apparent in the SPINE-vectors. In other words, it means that the vector features are not able to solve the task. In such cases, the selection of the most important features may prove inefficient and unfaithful such as in emotion detection.

- *Choice of mapping functions*: The divergence among the selection methods was less noticeable for feature selection than for descriptor selection. This can be explained by the high correlation observed among descriptors (Figure D.8), in contrast to features, which demonstrate weak correlation based on Pearson correlation (Figure 10.6b). This correlation might be handled in different ways by the mapping functions. Indeed, the utilization of different mapping functions strengthens confidence in feature selection. However, it is essential to choose these mapping functions judiciously, ensuring compatibility with the relationship between the features in question.

- *Lack of description interpretability*: While this explainability method serves as a valuable tool for phoneticians to understand the encoding of specific vocal characteristics, it is important to note that the provided description in this work lack interpretability, where it requires some phonetic expertise to be interpreted. This aspect of interpretability remains a subject for future exploration.

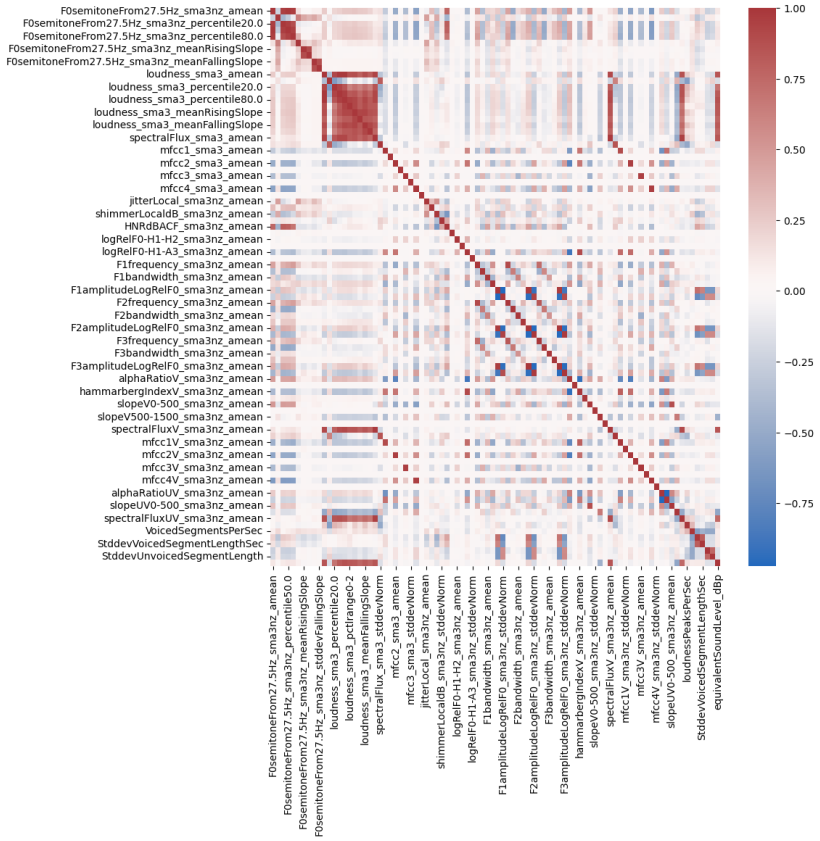


Figure D.8: Pearson correlation between OpenSmile eGeMAPs descriptors

Appendix E

Extra results for BA-LR calibration

Table E.1: Comparison of speaker recognition performance of BA-LR Speech-based and DNA-inspired versions on Test sets before and after fusion

Device	Sessions	BA-LR		Selective Fusion			
		EER (205 BAs)		EER		#BAs	
		DNA-inspired	Speech-based	DNA-inspired	Speech-based	DNA-inspired	Speech-based
1	1&2	1.21%	1.037%	1.81%	1.87%	123	132
	5&6	0.93%	0.96%	1.3%	1.2%	139	139
	3&4	0.78%	1.22%	1.82%	1.83%	146	149
	7&8	0.43%	0.43%	0.56%	0.5%	159	159
4	1&2	2.46%	2.07%	2.37%	2.37%	120	119
	3&4	6.85%	4.27%	2.79%	2.82%	164	144
5	1&2	11.39%	10.05%	7.45%	7.31%	105	101
	5&6	13.12	11.2%	7.89%	7.84%	113	128
	3&4	11.7%	10.72%	7.06%	7.18%	131	127
	7&8	13.85%	12.61%	7.48%	7.59%	123	124

Table E.2: Comparison of $\text{Cllr}_{min/act}$ between DNA-inspired and Speech-based versions of BA-LR before (Non-Calibrated) and after (Calibrated) applying calibration and fusion approaches

Device	Sessions	Non-Calibrated		Calibrated			
				Global		Fusion	
		DNA-inspired	Speech-based	DNA-inspired	Speech-based	DNA-inspired	Speech-based
1	1&2	0.03/0.2	0.037/0.60	0.03/0.06	0.037/0.081	0.07/0.11	0.07/0.10
	5&6	0.04/0.23	0.04/0.64	0.04/0.05	0.04/0.059	0.058/0.08	0.054/0.078
	3&4	0.029/0.28	0.042/0.64	0.029/0.049	0.042/0.065	0.07/0.093	0.07/0.08
	7&8	0.018/0.22	0.014/0.59	0.018/0.028	0.014/0.03	0.021/0.028	0.019/0.024
4	1&2	0.09/1.3	0.082/1.71	0.09/0.10	0.08/0.10	0.094/0.11	0.096/0.10
	3&4	0.24/8.38	0.16/8.26	0.24/0.25	0.16/0.16	0.10/0.13	0.1/0.12
5	1&2	0.4/6.03	0.35/8.78	0.4/0.42	0.36/0.38	0.27/0.30	0.26/0.3
	5&6	0.46/7.52	0.41/10.2	0.46/0.49	0.41/0.45	0.29/0.31	0.28/0.30
	3&4	0.38/7.25	0.35/10.0	0.38/0.40	0.35/0.38	0.26/0.27	0.26/0.27
	7&8	0.46/7.53	0.42/10.1	0.46/0.48	0.42/0.43	0.27/0.28	0.27/0.28

Bibliography

[1] John H.L. Hansen and Taufiq Hasan. "Speaker Recognition by Machines and Humans". In: *IEEE signal processing magasine*. 2015.

[2] Varun Sharma and P K Bansal. "A Review On Speaker Recognition Approaches And Challenges". In: *International Journal of Engineering Research & Technology (IJERT)*. 2013.

[3] Figen ERTAŞ. "FUNDAMENTALS OF SPEAKER RECOGNITION". In: *Journal of Engineering Sciences 2000*. 2000.

[4] Petr Cerva et al. "Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives". In: *Speech Communication* 55.10 (2013), pp. 1033–1046. DOI: https://doi.org/10.1016/j.specom.2013.06.017.

[5] Wiebke Toussaint Hutiri and Aaron Yi Ding. "Bias in Automated Speaker Recognition". In: *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.

[6] Gianni Fenu et al. "Fair Voice Biometrics: Impact of Demographic Imbalance on Group Fairness in Speaker Recognition". In: *Interspeech2021*. 2021.

[7] Gianni Fenu et al. "Improving Fairness in Speaker Recognition". In: *ESSE 2020: 2020 European Symposium on Software Engineering*. 2020.

[8] Sophie Noiret, Jennifer Lumetzberger, and Martin Kampel. "Bias and Fairness in Computer Vision Applications of the Criminal Justice System". In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2021.

[9] Julia Angwin et al. "Machine Bias". In: *ProPublica*. 2016.

[10] Kristian Lum and William Isaac. "To predict and serve?" In: *Significance*. 2016.

[11] Gianni Fenu, Hicham Lafhouli, and Mirko Marras. "Exploring Algorithmic Fairness in Deep Speaker Verification". In: *International Conference on Computational Science and Its Applications*. 2020.

[12] Varun Sharma and P K Bansal. "Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition." In: *ENFSI Forensic Speech and Audio Analysis Working Group Meeting*. 2015.

[13] Imen Ben-Amor and Jean-François Bonastre. "BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison". In: *International Workshop on Biometrics and Forensics (IWBF)*. 2022.

[14] S. Pruzansky and M. V. Mathews. "Talker-recognition procedure based on analysis of variance". In: *The journal of the acoustical society of America*. 1964.

[15] Douglas A. Reynolds Thomas F. Quatieri and Robert B. Dunn. “Speaker Verification Using Adapted Gaussian Mixture Models”. In: *IEEE Signal Processing Letters*. 2006.

[16] Najim Dehak et al. “Front-End Factor Analysis for Speaker Verification”. In: *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*. 2010.

[17] F. Soong et al. “A vector quantization approach to speaker recognition”. In: *ICASSP ’85. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 10. 1985, pp. 387–390. DOI: 10.1109/ICASSP.1985.1168412.

[18] D. Burton. “Text-dependent speaker verification using vector quantization source coding”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1987), pp. 133–143. DOI: 10.1109/TASSP.1987.1165110.

[19] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. USA: Kluwer Academic Publishers, 1991. ISBN: 0792391810.

[20] Tomi Kinnunen and Haizhou Li. “An overview of text-independent speaker recognition: From features to supervectors”. In: *Speech communication*. 2010.

[21] D. Reynolds. “Comparison of background normalization methods for text-independent speaker verification”. In: *Eurospeech 1997*. 1997.

[22] W.M. Campbell, D.E. Sturim, and D.A. Reynolds. “Support vector machines using GMM supervectors for speaker verification”. In: *IEEE Signal Processing Letters*. 2006.

[23] Patrick Kenny et al. “Joint factor analysis versus eigenchannels in speaker recognition”. In: *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*. 2007.

[24] Đorđe Grozdić et al. “Comparison of GMM/UBM and i-vector based speaker recognition systems”. In: Oct. 2015.

[25] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. “A statistical model-based voice activity detection”. In: *IEEE Signal Processing Letters*. 1999.

[26] Roberto Togneri and Daniel Pullella. “An Overview of Speaker Identification: Accuracy and Robustness Issues”. In: *IEEE Circuits and Systems Magazine*. 2011.

[27] Bishnu S Atal. “The History of Linear Prediction”. In: *IEEE Signal Processing Magazine*. 2006.

[28] Hermansky H. “Perceptual linear predictive (PLP) analysis of speech”. In: *J Acoust Soc Am*. 1990.

[29] Steven B. Davis and Paul Mermelstein. “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. In: *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*. 1980.

[30] S. Furui. “Cepstral analysis technique for automatic speaker verification”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1981.

[31] H. Hermansky and N. Morgan. “RASTA processing of speech”. In: *IEEE Trans. Speech Audio Processing*. 1994.

[32] Abdel rahman Mohamed, George E. Dahl, and Geoffrey Hinton. “Acoustic Modeling Using Deep Belief Networks”. In: *IEEE Transactions on Audio, Speech, and Language Processing*. 2012.

[33] Alec Radford et al. “Robust Speech Recognition via Large-Scale Weak Supervision”. In: *arXiv preprint*. 2022.

[34] Steffen Schneider et al. “WAV2VEC: UNSUPERVISED PRE-TRAINING FOR SPEECH RECOGNITION”. In: *Interspeech*. 2019.

[35] Alexei Baevski et al. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Interspeech*. 2021.

[36] Herve Jégou, Matthijs Douze, and Cordelia Schmid. “Product Quantization for Nearest Neighbor Search”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010.

[37] Alec Radford et al. “Improving language understanding with unsupervised learning”. In: *Technical report, OpenAI*. 2018.

[38] Sanyuan Chen et. al. “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”. In: *IEEE Journal of Selected Topics in Signal Processing*. 2021.

[39] Gang Liu et al. “The Microsoft System for VoxCeleb Speaker Recognition Challenge 2022”. In: *arXiv:2209.11266*. 2022.

[40] Shu wen Yang et al. *SUPERB: Speech processing Universal PERformance Benchmark*. 2021. arXiv: 2105.01051 [cs.CL].

[41] Ehsan Variani et al. “DEEP NEURAL NETWORKS FOR SMALL FOOTPRINT TEXT-DEPENDENT SPEAKER VERIFICATION”. In: *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. 2014.

[42] Ian J. Goodfellow et al. “Maxout Networks”. In: *Proceedings of the 30th International Conference on Machine Learning, PMLR*. 2013.

[43] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. “A time delay neural network architecture for efficient modeling of long temporal contexts”. In: *Interspeech*. 2015.

[44] A. Waibel et al. “Phoneme recognition using time-delay neural networks”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1989.

[45] David Snyder, Daniel Garcia-Romero, and Daniel Povey. “TIME DELAY DEEP NEURAL NETWORK-BASED UNIVERSAL BACKGROUND MODELS FOR SPEAKER RECOGNITION”. In: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 2015.

[46] David Snyder et al. “Deep Neural Network Embeddings for Text-Independent Speaker Verification”. In: *Interspeech 2017*. 2017.

[47] David Snyder et al. “X-vectors: Robust DNN embeddings for speaker recognition”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*. 2018.

[48] David Snyder et al. “SPEAKER RECOGNITION FOR MULTI-SPEAKER CONVERSATIONS USING X-VECTORS”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*. 2019.

[49] Jesus Villalba et. al. “State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18”. In: *Interspeech*. 2019.

[50] David Snyder et al. “The JHU Speaker Recognition System for the VOiCES 2019 Challenge”. In: *Interspeech*. 2019.

[51] Daniel Povey et al. “Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks”. In: *Interspeech*. 2018.

[52] Qian-Bei Hong et al. “Statistics Pooling Time Delay Neural Network Based on X-Vector for Speaker Verification”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020.

[53] Boji Liu et al. “Time Delay Recurrent Neural Network for Speech Recognition”. In: *3rd International Conference on Machine Vision and Information Technology (CMVIT 2019)*. 2019.

[54] Ying Qin et al. “Automatic speech assessment for people with aphasia using TDNN-BLSTM with multi-task learning”. In: *Interspeech*. 2018.

[55] Xiaoxiao Miao, Ian McLoughlin, and Yonghong Yan. “A New Time-Frequency Attention Mechanism for TDNN and CNN-LSTM-TDNN, with Application to Language Identification”. In: *Interspeech*. 2019.

[56] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[57] Hossein Zeinali et al. “BUT System Description to VoxCeleb Speaker Recognition Challenge 2019”. In: *arXiv:1910.12592*. 2019.

[58] Ya-Qi Yu, Lei Fan, and Wu-Jun Li. “ENSEMBLE ADDITIVE MARGIN SOFTMAX FOR SPEAKER VERIFICATION”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[59] Mohammad MohammadAmini et al. “Learning noise robust ResNet-based speaker embedding for speaker recognition”. In: *The Speaker and Language Recognition Workshop (Odyssey)*. 2022.

[60] Daniel Garcia-Romero, Gregory Sell, and Alan McCree. “MagNetO: X-vector Magnitude Estimation Network plus Offset for Improved Speaker Recognition”. In: *The Speaker and Language Recognition Workshop (Odyssey 2020)*. 2020.

[61] Insoo Kim et al. “Deep Speaker Representation Using Orthogonal Decomposition and Recombination for Speaker Verification”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[62] Zhiming Wang et al. “MULTI-RESOLUTION MULTI-HEAD ATTENTION IN DEEP SPEAKER EMBEDDING”. In: *ICASSP*. 2020.

[63] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. “VoxCeleb2: Deep Speaker Recognition”. In: *Interspeech*. 2018.

[64] Chao Li et. al. “Deep Speaker: an End-to-End Neural Speaker Embedding System”. In: *arXiv preprint arXiv:1705.02304*. 2017.

[65] Sarthak Yadav and Atul Rai. “FREQUENCY AND TEMPORAL CONVOLUTIONAL ATTENTION FOR TEXT-INDEPENDENT SPEAKER RECOGNITION”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020.

[66] Weidi Xie et al. “UTTERANCE-LEVEL AGGREGATION FOR SPEAKER RECOGNITION IN THE WILD”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[67] Tianyan Zhou et al. “CNN WITH PHONETIC ATTENTION FOR TEXT-INDEPENDENT SPEAKER VERIFICATION”. In: *ASRU*. 2019.

[68] Jianfeng Zhou et al. “Deep Speaker Embedding Extraction with Channel-Wise Feature Responses and Additive Supervision Softmax Loss Function”. In: *Interspeech*. 2019.

[69] Magdalena Rybicka et al. “Spine2Net: SpineNet with Res2Net and Time-Squeeze-and-Excitation Blocks for Speaker Recognition”. In: *INTERSPEECH 2021*. 2021.

[70] Mickael Rouvier and Pierre-Michel Bousquet. “STUDYING SQUEEZE-AND-EXCITATION USED IN CNN FOR SPEAKER VERIFICATION”. In: *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2021.

[71] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification”. In: *Proc. Interspeech*. 2020.

[72] Yuan Lei et al. “ZXIC Speaker Verification System for FFSVC 2022 Challenge”. In: *The 2022 Far-field Speaker Verification Challenge(FFSVC2022)*. 2022.

[73] Yang Zhang et al. *MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification*. 2022. arXiv: 2203.15249 [cs.SD].

[74] Sergey Novoselov et al. "Robust Speaker Recognition with Transformers Using wav2vec 2.0". In: *ArXiv* abs/2203.15095 (2022). URL: https://api.semanticscholar.org/CorpusID:247778411.

[75] Arun Babu et al. *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. 2021. arXiv: 2111.09296 [cs.CL].

[76] Mickael Rouvier, Pierre-Michel Bousquet, and Jarod Duret. "Study on the temporal pooling used in deep neural networks for speaker verification". In: *29th European Signal Processing Conference (EUSIPCO)*. 2021.

[77] Shuai Wang et al. "Revisiting the Statistics Pooling Layer in Deep Speaker Embedding Learning". In: *12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. 2021.

[78] Gautam Bhattacharya, Jahangir Alam, and Patrick Kenny. "Deep Speaker Embeddings for Short-Duration Speaker Verification". In: *Proc. Interspeech 2017*. 2017, pp. 1517–1521. DOI: 10.21437/Interspeech.2017-1575.

[79] F A Rezaur rahman Chowdhury et al. "Attention-Based Models for Text-Dependent Speaker Verification". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 5359–5363. DOI: 10.1109/ICASSP.2018.8461587.

[80] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding". In: *Interspeech*. 2018.

[81] Miquel India, Pooyan Safari, and Javier Hernando. "Self Multi-Head Attention for Speaker Recognition". In: *Interspeech*. 2019.

[82] Bin Gu and Wu Guo. *An Improved Deep Neural Network for Modeling Speaker Characteristics at Different Temporal Scales*. 2020. arXiv: 2001.04584 [eess.AS].

[83] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2022), pp. 5962–5979. DOI: 10.1109/tpami.2021.3087709. URL: https://doi.org/10.1109%2Ftpami.2021.3087709.

[84] Weiyang Liu et al. *SphereFace: Deep Hypersphere Embedding for Face Recognition*. 2018. arXiv: 1704.08063 [cs.CV].

[85] Zili Huang, Shuai Wang, and Kai Yu. "Angular Softmax for Short-Duration Text-independent Speaker Verification". In: *Proc. Interspeech 2018*. 2018, pp. 3623–3627. DOI: 10.21437/Interspeech.2018-1545.

[86] Xu Xiang et al. *Margin Matters: Towards More Discriminative Deep Neural Network Embeddings for Speaker Recognition*. 2019. arXiv: 1906.07317 [eess.AS].

[87] Zhongxin Bai and Xiao-Lei Zhang. “Speaker recognition based on deep learning: An overview”. In: *Elsevier*. 2021.

[88] Qin Jin and Alexander H. Waibel. “Application of LDA to speaker recognition”. In: *Interspeech*. 2000. URL: https://api.semanticscholar.org/CorpusID:326002.

[89] Keiji Miura. “An Introduction to Maximum Likelihood Estimation and Information Geometry”. In: *Interdisciplinary Information Sciences (IIS)* 17 (2011). DOI: 10.4036/iis.2011.155.

[90] Patrick Kenny. “Bayesian Speaker Verification with Heavy-Tailed Priors”. In: *The Speaker and Language Recognition Workshop*. 2010.

[91] Simon J.D. Prince and James H. Elder. “Probabilistic Linear Discriminant Analysis for Inferences About Identity”. In: *2007 IEEE 11th International Conference on Computer Vision*. 2007. DOI: 10.1109/ICCV.2007.4409052.

[92] T.K. Moon. “The expectation-maximization algorithm”. In: *IEEE Signal Processing Magazine* (1996). DOI: 10.1109/79.543975.

[93] Najim Dehak et al. “Front-End Factor Analysis for Speaker Verification”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798. DOI: 10.1109/TASL.2010.2064307.

[94] Zhiyuan Peng et al. “Unifying Cosine and PLDA Back-ends for Speaker Verification”. In: *Proc. Interspeech 2022*. 2022, pp. 336–340. DOI: 10.21437/Interspeech.2022-10021.

[95] David A. van Leeuwen and Niko Brümmer. “An Introduction to Application-Independent Evaluation of Speaker Recognition Systems”. In: *Speaker Classification I: Fundamentals, Features, and Methods*. Ed. by Christian Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 330–353. URL: https://doi.org/10.1007/978-3-540-74200-5_19.

[96] A. Martin et al. “The det curve in assessment of detection task performance”. In: *The DET Curve in Assessment of Detection Task Performance* (Jan. 1997), pp. 1895–1898.

[97] Andrzej Drygajlo. “Automatic Speaker Recognition for Forensic Case Assessment and Interpretation”. In: 2012. URL: https://api.semanticscholar.org/CorpusID:11073685.

[98] Jean-François Bonastre et al. “Forensic speaker recognition”. In: *IEEE* (Mar. 2010).

[99] Christophe Champod and Didier Meuwly. “Inference of identity in forensic speaker recognition”. In: *Speech Communication* (2000), pp. 193–203.

[100] Andrzej Drygajlo et al. “Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises”. In: 2016. URL: https://api.semanticscholar.org/CorpusID:52363149.

[101] Anil Alexander. “Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions”. In: *Forensic Linguistics* (Jan. 2007). DOI: 10.5075/epfl-thesis-3367.

[102] Mohammed Algabri et al. “Automatic Speaker Recognition for Mobile Forensic Applications”. In: *Mob. Inf. Syst.* 2017 (2017), 6986391:1–6986391:6. URL: https://api.semanticscholar.org/CorpusID:27633228.

[103] Finnian Kelly et al. “Deep Neural Network Based Forensic Automatic Speaker Recognition in VOCALISE using x-Vectors”. In: *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*. 2019. URL: http://www.aes.org/e-lib/browse.cfm?elib=20477.

[104] Francesco Sigona and Mirko Grimaldi. *Validation of an ECAPA-TDNN system for Forensic Automatic Speaker Recognition under case work conditions*. 2023. arXiv: 2305.10805 [cs.SD].

[105] Dávid Sztahó and Attila Fejes. “Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings”. In: *Journal of Forensic Sciences* 68.3 (2023), pp. 871–883. DOI: https://doi.org/10.1111/1556-4029.15250. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.15250.

[106] Battineni Venkata Kishore Babu et al. “Forensic Speaker Recognition System using Machine Learning”. In: *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. 2023. DOI: 10.1109/ICSCDS56580.2023.10104687.

[107] Emmanuel Maqueda et al. “Triplet loss-based embeddings for forensic speaker identification in Spanish”. In: *Neural Computing and Applications* (2021). DOI: 10.1007/s00521-021-06408-6. URL: https://doi.org/10.1007%2Fs00521-021-06408-6.

[108] Sajid Saleem et al. “Forensic speaker recognition: A new method based on extracting accent and language information from short utterances”. In: *Forensic Science International: Digital Investigation* 34 (2020), p. 300982. ISSN: 2666-2817. DOI: https://doi.org/10.1016/j.fsidi.2020.300982. URL: https://www.sciencedirect.com/science/article/pii/S2666281720300500.

[109] Anil Alexander et al. “VOCALISE : A forensic automatic speaker recognition system supporting spectral , phonetic , and user-provided features”. In: 2016. URL: https://api.semanticscholar.org/CorpusID:202714051.

[110] Michael Jessen et al. “Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)”. In: *Speech Communication* 111 (May 2019). DOI: 10.1016/j.specom.2019.05.002.

[111] Bernard Robertson, G.A. Vignaux, and Charles Berger. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. 1995.

[112] C G G Aitken and D.Lucy. “Evaluation of trace evidence in the form of multivariate data”. In: *Journal of the royal statistical society* (2004).

[113] Ian W. Evett. “Towards a uniform framework for reporting opinions in forensic science casework”. In: *Science & Justice* 38 (1998), pp. 198–202. URL: https://api.semanticscholar.org/CorpusID:111282083.

[114] Phil Rose. “Technical forensic speaker recognition: Evaluation, types and testing of evidence”. In: *Computer speech and language* (2006).

[115] Colin Aitken and Franco Taroni. *Statistics and the evaluation of evidence for forensic scientists*. English. 2nd edition. United States: John Wiley & Sons Inc., 2004. ISBN: 978-0-470-84367-3. DOI: 10.1002/0470011238.

[116] Philip rose. “Forensic Speaker Identification”. In: *Taylor and Francis* (2002).

[117] Philip Rose and Geoffrey Stewart Morrison. “A response to the UK Position Statement on forensic speaker comparison”. In: *International Journal of Speech, Language and the Law* (2009), 139–163.

[118] Geoffrey Stewart Morrison et al. “Statistical Models in Forensic Voice Comparison”. In: *Handbook of forensic statistics* (2020), pp. 451–479.

[119] Daniel Ramos et al. “From Biometric Scores to Forensic Likelihood Ratios”. In: Feb. 2017. ISBN: 978-3-319-50671-5.

[120] Daniel Ramos. “Forensic evaluation of the evidence using automatic speaker recognition systems”. In: 2014. URL: https://api.semanticscholar.org/CorpusID:191937064.

[121] Annabel Bolck, Haifang Ni, and Martin Lopatka. “Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison”. In: *Journal of Law, Probability and Risk* (2015), pp. 243–266.

[122] Xiao-Hong Chen et al. “Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features”. In: *Forensic science international* (2018), pp. 101–110.

[123] Anna Jeannette Leegwater et al. “Performance study of a score-based likelihood ratio system for forensic fingermark comparison”. In: *Journal of forensic sciences* (2017).

[124] Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez. "Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors". In: *Odyssey* (2016).

[125] Michael Carne and Shunich Ishihara. "Feature-Based Forensic Text Comparison Using a Poisson Model for Likelihood Ratio Estimation". In: *ACL Anthology* (2020), pp. 32–42.

[126] Joaquin Gonzalez-Rodriguez et al. "Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems". In: *Forensic Science International* 155.2 (2005), pp. 126–140. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2004.11.007. URL: https://www.sciencedirect.com/science/article/pii/S0379073804007509.

[127] Niko Brümmer and Johan A. du Preez. "Application-independent evaluation of speaker detection". In: *Computer Speech and Language*. 2006.

[128] Joaquín González-Rodríguez, Javier Ortega-Garcia, and Jose Juan Lucena-Molina. "On the application of the Bayesian approach in real forensic conditions with GMM-based systems". In: *The Speaker and Language Recognition Workshop*. 2001. URL: https://api.semanticscholar.org/CorpusID:43653017.

[129] Daniel Ramos and Joaquin Gonzalez-Rodriguez. "Reliable support: Measuring calibration of likelihood ratios". In: *Forensic Science International* 230.1 (2013). EAFS 2012 6th European Academy of Forensic Science Conference The Hague, 20-24 August 2012, pp. 156–169. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2013.04.014. URL: https://www.sciencedirect.com/science/article/pii/S0379073813002375.

[130] Andrew van Es et al. "Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis". In: *Science & Justice* 57.3 (2017), pp. 181–192. ISSN: 1355-0306. DOI: https://doi.org/10.1016/j.scijus.2017.03.002. URL: https://www.sciencedirect.com/science/article/pii/S1355030617300266.

[131] David A. van Leeuwen and Niko Brümmer. *The distribution of calibrated likelihood-ratios in speaker recognition*. 2013. arXiv: 1304.1199 [stat.AP].

[132] Joaquin Gonzalez-Rodriguez et al. "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.7 (2007), pp. 2104–2115.

[133] Geoffrey Stewart Morrison. "Measuring the validity and reliability of forensic likelihood-ratio systems". In: *Science & Justice* 51.3 (2011), pp. 91–98. ISSN: 1355-0306. DOI: https://doi.org/10.1016/j.scijus.2011.03.002. URL: https://www.sciencedirect.com/science/article/pii/S1355030611000256.

[134] David van der Vloed and Tina Cambier-Langevel. "How we use automatic speaker comparison in forensic practice." In: *International Journal of Speech, Language & the Law* (2022).

[135] Brandon L. Garrett and Cynthia Rudin. "The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice". In: *Cornell Law Review, Forthcoming, Duke Law School Public Law & Legal Theory Series No. 2023-03* (2023).

[136] Brandon L. Garrett and Cynthia Rudin. "Interpretable algorithmic forensics". In: *Proceedings of the National Academy of Sciences* 120.41 (2023), e2301842120. DOI: 10.1073/pnas.2301842120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2301842120.

[137] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[138] Lauren Kirchner. "New York City Moves to Create Accountability for Algorithms". In: *ProPublica* (2017).

[139] Th. Kirat et al. "Fairness and explainability in automatic decision-making systems. A challenge for computer science and law". In: *EURO Journal on Decision Processes* 11 (2023), p. 100036.

[140] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation". In: *Forensic Science International* 276 (2017), pp. 142–153. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2016.03.048. URL: https://www.sciencedirect.com/science/article/pii/S0379073816301359.

[141] Rolf J.F. Ypma, Daniel Ramos, and Didier Meuwly. "AI-based Forensic Evaluation in Court: The Desirability of Explanation and the Necessity of Validation". English. In: *Artificial Intelligence (AI) in Forensic Sciences*. Ed. by Zeno Geradts and Katrin Franke. United States: Wiley, Aug. 2023. ISBN: 978-1-119-81332-3.

[142] Abiodun A. Solanke. "Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models". In: *Forensic Science International: Digital Investigation* 42 (2022). Proceedings of the Twenty-Second Annual DFRWS USA, p. 301403. ISSN: 2666-2817. DOI: https://doi.org/10.1016/j.fsidi.2022.301403. URL: https://www.sciencedirect.com/science/article/pii/S2666281722000841.

[143] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. "Explanation in AI and law: Past, present and future". In: *Artificial Intelligence* 289 (2020), p. 103387. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2020.103387. URL: https://www.sciencedirect.com/science/article/pii/S0004370220301375.

[144] Adrien Bibal et al. *Impact of Legal Requirements on Explainability in Machine Learning*. 2020. arXiv: 2007.05479 [cs.AI].

[145] Cynthia Rudin. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. 2019. arXiv: 1811.10154 [stat.ML].

[146] P. Jonathon Phillips and Mark Przybocki. *Four Principles of Explainable AI as Applied to Biometrics and Facial Forensic Algorithms*. 2020. arXiv: 2002.01014 [cs.CV].

[147] Ashley S. Deeks. "The Judicial Demand for Explainable Artificial Intelligence". In: *Law & Society: Public Law - Courts eJournal* (2019).

[148] Michael M. O'Hear. "Appellate Review of Sentence Explanations: Learning from the Wisconsin and Federal Experiences". In: *Marquette Law Review, Issue 2 Symposium: Criminal Appeals: Past, Present, and Future* (2019).

[149] Sajid Ali et al. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". In: *Information Fusion* 99 (2023), p. 101805. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2023.101805. URL: https://www.sciencedirect.com/science/article/pii/S1566253523001148.

[150] Johanna Moore and William Swartout. "Explanation in expert systems: A survey". In: (Jan. 1989).

[151] Committee on Technology National Science, Technology Council, and Penny Hill Press. "Preparing for the Future of Artificial Intelligence". In: (2016).

[152] Dave Gunning et al. "Special Issue:DARPA's Explainable Artificial Intelligence (XAI) Program". In: *Applied AI letters*. 2021.

[153] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation". In: *International Data Privacy Law* (2017), pp. 76–99. eprint: https://academic.oup.com/idpl/article-pdf/7/2/76/17932196/ipx005.pdf. URL: https://doi.org/10.1093/idpl/ipx005.

[154] Jarek Gryz and Marcin Rojszczak. "Black box algorithms and the rights of individuals: No easy solution to the "explainability" problem". In: *Internet Policy Review* 10 (June 2021). DOI: 10.14763/2021.2.1564.

[155] Bryan Casey, Ashkon Farhangi, and Roland Vogl. "RETHINKING EXPLAINABLE MACHINES :THE GDPR'S "RIGHT TO EXPLANATION" DEBATE AND THE RISE OF A LGORITHMIC AUDITS IN ENTERPRISE". In: *BERKELEY TECHNOLOGY LAW JOURNAL* 10 (2019).

[156] European Union Artificial Intelligence Act. "European Commission". In: (2021). URL: https://www.euaiact.com/.

[157] Andreas Holzinger et al. *What do we need to build explainable AI systems for the medical domain?* 2017. arXiv: 1712.09923 [cs.AI].

[158] Rich Caruana et al. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2015. ISBN: 9781450336642. DOI: 10.1145/2783258.2788613. URL: https://doi.org/10.1145/2783258.2788613.

[159] Sondes Abderrazek et al. "Interpreting Deep Representations of Phonetic Features via Neuro-Based Concept Detector: Application to Speech Disorders Due to Head and Neck Cancer". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[160] Dan Simmons. "BBC fools HSBC voice recognition security system". In: *BBC* (2017). URL: https://www.bbc.com/news/technology-39965545?ref=hackernoon.com.

[161] Chaofan Chen et al. "A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations". In: *Decision Support Systems* 152 (2022), p. 113647.

[162] Petter Eilif de Lange et al. "Explainable AI for Credit Assessment in Banks". In: *Journal of Risk and Financial Management* 15.12 (2022).

[163] Dylan Tokar. "Google Cloud Launches Anti-Money-Laundering Tool for Banks, Betting on the Power of AI". In: *Risk & Compliance Journal* (2023).

[164] Pidikiti Supriya et al. "Loan prediction by using machine learning models". In: *International Journal of Engineering and Techniques* 5.2 (2019), pp. 144–147.

[165] Xu Zhu et al. "Explainable prediction of loan default based on machine learning models". In: *Data Science and Management* 6.3 (2023), pp. 123–133. ISSN: 2666-7649. DOI: https://doi.org/10.1016/j.dsm.2023.04.003. URL: https://www.sciencedirect.com/science/article/pii/S2666764923000218.

[166] Wexler R. "When a Computer Program Keeps You in Jail: How Computers are Harming Criminal Justice". In: (2017).

[167] John Lightbourne. "Damned Lies & Criminal Sentencing Using Evidence-Based Tools". In: (May 2017).

[168] Andrew Bell et al. "It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 248–266. ISBN: 9781450393522. DOI: 10.1145/3531146.3533090. URL: https://doi.org/10.1145/3531146.3533090.

[169] Chuck Howell. "A framework for addressing fairness in consequential machine learning". In: 2018. URL: https://api.semanticscholar.org/CorpusID:211201589.

[170] Caroline Wang et al. *In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction*. 2022. arXiv: 2005.04176 [stat.ML].

[171] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[172] Doshi-Velez Finale, and Mason Kortz. "Accountability of AI Under the Law: The Role of Explanation". In: *Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper* (2017).

[173] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38. DOI: https://doi.org/10.1016/j.artint.2018.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[174] Adrian Erasmus, Tyler Brunet, and · Fisher. "What is Interpretability?" In: *Philosophy & Technology* Online (Dec. 2021). DOI: 10.1007/s13347-020-00435-2.

[175] David A. Broniatowski. "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence". In: 2021. URL: https://api.semanticscholar.org/CorpusID:234892700.

[176] Diogo Vieira Carvalho, Eduardo Marques Pereira, and Jaime S. Cardoso. "Machine Learning Interpretability: A Survey on Methods and Metrics". In: *Electronics* (2019). URL: https://api.semanticscholar.org/CorpusID:199659548.

[177] Zachary Chase Lipton. "The Mythos of Model Interpretability". In: *CoRR* abs/1606.03490 (2018). arXiv: 1606.03490. URL: http://arxiv.org/abs/1606.03490.

[178] Ricards Marcinkevics and Julia E. Vogt. "Interpretability and Explainability: A Machine Learning Zoo Mini-tour". In: *CoRR* abs/2012.01805 (2020). URL: https://arxiv.org/abs/2012.01805.

[179] Krzysztof Fiok et al. "Explainable artificial intelligence for education and training". In: *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 19 (July 2021), p. 154851292110286. DOI: 10.1177/15485129211028651.

[180] Peter ED Love et al. *Explainable Artificial Intelligence: Precepts, Methods, and Opportunities for Research in Construction*. 2023. arXiv: 2211.06579 [cs.AI].

[181] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[182] Leilani H. Gilpin et al. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. 2019. arXiv: 1806.00069 [cs.AI].

[183] Christoph Molnar. “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable”. In: (2019).

[184] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. “Explainable AI: A Review of Machine Learning Interpretability Methods”. In: *Entropy* 23.1 (2021). ISSN: 1099-4300. DOI: 10.3390/e23010018. URL: https://www.mdpi.com/1099-4300/23/1/18.

[185] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. “Methods for interpreting and understanding deep neural networks”. In: *Digital Signal Processing* 73 (2018), pp. 1–15. ISSN: 1051-2004. DOI: https://doi.org/10.1016/j.dsp.2017.10.011. URL: https://www.sciencedirect.com/science/article/pii/S1051200417302385.

[186] Andreas Holzinger et al. “Causability and explainabilty of artificial intelligence in medicine”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9 (Apr. 2019), e1312. DOI: 10.1002/widm.1312.

[187] Ivars NAMATEVS, Kaspars SUDARS, and Artis DOBRAJS. “Interpretability versus Explainability: Classification for Understanding Deep Learning Systems and Models”. In: *Computer Assisted Methods in Engineering and Science* (2022).

[188] Timo Speith. “A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods”. In: June 2022, pp. 2239–2250. DOI: 10.1145/3531146.3534639.

[189] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. “Explainable Deep Learning Models in Medical Image Analysis”. In: *Journal of Imaging* 6 (June 2020), p. 52. DOI: 10.3390/jimaging6060052.

[190] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. *Explainable AI: current status and future directions*. 2021. arXiv: 2107.07045 [cs.LG].

[191] Johannes Fürnkranz. “Rule Set”. In: (Jan. 2017), pp. 1121–1121.

[192] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. arXiv: 1312.6034 [cs.CV].

[193] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG].

[194] Jost Tobias Springenberg et al. *Striving for Simplicity: The All Convolutional Net*. 2015. arXiv: 1412.6806 [cs.LG].

[195] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[196] Sebastian Lapuschkin et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLoS ONE* 10 (July 2015), e0130140. DOI: 10.1371/journal.pone.0130140.

[197] L Shapley. "Quota solutions op n-person games1". In: *Edited by Emil Artin and Marston Morse* (1953), p. 343.

[198] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. arXiv: 1602.04938 [cs.LG].

[199] Scott Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: (2017). arXiv: 1705.07874 [cs.AI].

[200] Kjell Hausken and Matthias Mohr. "The value of a player in n-person games". In: *Social Choice and Welfare* 18 (Aug. 2001), pp. 465–483. DOI: 10.1007/s003550000070.

[201] Hannah Muckenhirn et al. "Understanding and Visualizing Raw Waveform-Based CNNs". In: *Interspeech*. 2019. URL: https://api.semanticscholar.org/CorpusID:197674956.

[202] Pengqi Li et al. "Reliable Visualization for Deep Speaker Recognition". In: *Interspeech* (2022).

[203] Homanga Bharadhwaj. "Layer-Wise Relevance Propagation for Explainable Deep Learning Based Speech Recognition". In: *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (2018).

[204] Ali Raza Syed and Michael I. Mandel. "Data Valuation for Acoustic Models in Automatic Speech Recognition". In: *NeurIPS 2020 Workshop on Dataset Curation and Security*. 2020.

[205] Karla Markert et al. "Visualizing Automatic Speech Recognition – Means for a Better Understanding?" In: *ISCA Symposium on Security and Privacy in Speech Communication*. 2022.

[206] Suk-Young Lim, Dong-Kyu Chae, and Sang-Chul Lee. "Detecting Deepfake Voice Using Explainable Deep Learning Techniques". In: *Applied Science* (2022).

[207] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLoS ONE* 10 (2015). URL: https://api.semanticscholar.org/CorpusID:9327892.

[208] Sören Becker et al. *Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals*. 2019. arXiv: 1807.03418 [cs.SD].

[209] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[210] Sunit Sivasankaran, Emmanuel Vincent, and Dominique Fohr. "Explaining deep learning models for speech enhancement". In: *Interspeech 2021*. 2021.

[211] B Budowle et al. "Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from vntr loci, for use in forensic comparisons". In: *American Journal of Human Genetics* (1991), pp. 841–855.

[212] J Brookfield. "The effect of population subdivision on estimates of the likelihood ratio in criminal-cases using single-locus DNA probes". In: *Heredity* (1992).

[213] A. Collins and N. E. morton. "Likelihood ratios for DNA identification". In: *Medical sciences* (1994).

[214] Anders Nordgaard and Birgitta Rasmusson. "The likelihood ratio as value of evidence—more than a question of numbers". In: *Law, probability and Risk* (2012).

[215] Steven P. Lund and Hari Iyer. "Likelihood Ratio as Weight of Forensic Evidence: A Closer Look". In: *Journal of Research of National Institute of Standards and Technology* (2017).

[216] 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". In: *Forensic Sci Int Genet* (2012), pp. 679–688.

[217] Anna G. Shestak et al. "Allelic Dropout Is a Common Phenomenon That Reduces the Diagnostic Yield of PCR-Based Sequencing of Targeted Gene Panels". In: *Frontiers in Genetics* 12 (2021).

[218] F Van Nieuwerburgh et al. "Impact of allelic dropout on evidential value of forensic DNA profiles using RMNE". In: *Oxford journal bioinformatics* (2010).

[219] Frederick R.Bieber et al. "Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion". In: *BMC genetics* (2016), pp. 679–688.

[220] Peter Gill et al. "An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA". In: *Forensic science international* 112.1 (2000), pp. 17–40.

[221] David J Balding and John Buckleton. "Interpreting low template DNA profile". In: *National library of medicine* (2010).

[222] Peter Gill, James Curran, and Cedric Neumann. “Interpretation of complex DNA profiles using Tippett plots”. In: *Forensic science international: Genetics supplement series 1* (2007), pp. 646–648.

[223] Torben Tvedebrink et al. “Estimating the probability of allelic drop-out of STR alleles in forensic genetics”. In: *Forensic Sci Int Genet* (2012), pp. 679–688.

[224] Adele A. Mitchell et al. “Likelihood ratio statistics for DNA mixtures allowing for drop-out and drop-in”. In: *Forensic Science International: Genetics Supplement Series* 3.1 (2011). Progress in Forensic Genetics 14, e240–e241. ISSN: 1875-1768. DOI: https://doi.org/10.1016/j.fsigss.2011.08.119. URL: https://www.sciencedirect.com/science/article/pii/S1875176811001211.

[225] Navin Chatlani and John J. Soraghan. “Local binary patterns for 1-D signal processing”. In: *2010 18th European Signal Processing Conference*. 2010.

[226] li Deng et al. “Binary coding of speech spectrograms using a deep auto-encoder”. In: Sept. 2010, pp. 1692–1695. DOI: 10.21437/Interspeech.2010-487.

[227] Jean-François Bonastre et al. “Speaker modeling using local binary decisions”. In: *Proc. Interspeech 2011*. 2011, pp. 13–16. DOI: 10.21437/Interspeech.2011-4.

[228] J.F. Bonastre et al. “Discriminant binary data representation for speaker recognition”. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011, pp. 5284–5287. DOI: 10.1109/ICASSP.2011.5947550.

[229] Gabriel Hernández-Sierra, Jean-François Bonastre, and José Lara. “Speaker Recognition Using a Binary Representation and Specificities Models”. In: Sept. 2012, pp. 732–739. ISBN: 978-3-642-33274-6.

[230] Petros Boufounos and Shantanu Rane. “Secure binary embeddings for privacy preserving nearest neighbors”. In: *2011 IEEE International Workshop on Information Forensics and Security*. 2011, pp. 1–6. DOI: 10.1109/WIFS.2011.6123149.

[231] José Portêlo et al. “Secure binary embeddings of front-end factor analysis for privacy preserving speaker verification”. In: *Proc. Interspeech 2013*. 2013, pp. 2494–2498. DOI: 10.21437/Interspeech.2013-417.

[232] José Portêlo et al. “Privacy-preserving speaker verification using secure binary embeddings”. In: *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2014, pp. 1268–1272. DOI: 10.1109/MIPRO.2014.6859762.

[233] Woojay Jeon and Yan-Ming Cheng. “Efficient speaker search over large populations using kernelized locality-sensitive hashing”. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012, pp. 4261–4264. DOI: 10.1109/ICASSP.2012.6288860.

[234] Brian Kulis and Kristen Grauman. "Kernelized locality-sensitive hashing for scalable image search". In: *2009 IEEE 12th International Conference on Computer Vision*. 2009, pp. 2130–2137. DOI: 10.1109/ICCV.2009.5459466.

[235] Lantian Li et al. "Binary speaker embedding". In: *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. 2016, pp. 1–4. DOI: 10.1109/ISCSLP.2016.7918381.

[236] Jiaying Wang et al. "Ordered and Binary Speaker Embedding". In: *Interspeech 2023* (2023).

[237] Charles Dugas et al. "Incorporating Second-Order Functional Knowledge for Better Option Pricing". In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000. URL: https://proceedings.neurips.cc/paper_files/paper/2000/file/44968aece94f667e4095002d140b5896-Paper.pdf.

[238] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "VoxCeleb: A Large-Scale Speaker Identification Dataset". In: *Interspeech*. 2017.

[239] Daniel Povey et al. "The Kaldi speech recognition toolkit". In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (Jan. 2011).

[240] Reginald Smith. "A mutual information approach to calculating nonlinearity". In: *Stat* (2015). DOI: 10.1002/sta4.96. URL: https://doi.org/10.1002%2Fsta4.96.

[241] Pierre-Michel Bousquet and Jean-François Bonastre. "Typicality extraction in a Speaker Binary Keys model". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012, pp. 1713–1716. DOI: 10.1109/ICASSP.2012.6288228.

[242] Geoffrey Stewart Morrison and Ewald Enzinger. "Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality". In: *Science & Justice* 58.1 (2018), pp. 47–58. ISSN: 1355-0306. DOI: https://doi.org/10.1016/j.scijus.2017.06.005. URL: https://www.sciencedirect.com/science/article/pii/S1355030617300849.

[243] Vincent Hughes, Ashley Brereton, and Erica Gold. "Reference sample size and the computation of numerical likelihood ratios using articulation rate". In: *York papers in Linguistics Series 2* (2013).

[244] Mitchell McLaren et al. "The speakers in the wild (sitw) speaker recognition database". In: *Interspeech* (2016).

[245] Mahesh Kumar Nandwana et al. "The VOiCES from a Distance Challenge 2019 Evaluation Plan". In: *Interspeech* (2019).

[246] Mahesh Kumar Nandwana et al. "The VOiCES from a Distance Challenge 2019: Analysis of Speaker Verification Results and Remaining Challenges". In: *Odyssey* (2020).

[247] Sanjoy Sarkar et al. “Accuracy and interpretability trade-offs in machine learning applied to safer gambling”. In: *CEUR Workshop Proceedings*. Vol. 1773. CEUR Workshop Proceedings. 2016.

[248] Shuai Wang, Yanmin Qian, and Kai Yu. “What Does the Speaker Embedding Encode?” In: *interspeech* (2017).

[249] Desh Raj et al. “PROBING THE INFORMATION ENCODED IN X-VECTORS”. In: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 2019.

[250] Zied Elloumi et al. “Analyzing Learned Representations of a Deep ASR Performance Prediction Model”. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018.

[251] Archiki Prasad and Preethi Jyothi. “How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems”. In: *58th Annual Meeting of the Association for Computational Linguistics*. 2020.

[252] Shammur A. Chowdhury et al. “What Does an End-to-End Dialect Identification Model Learn About Non-Dialectal Information?” In: *Proc. INTERSPEECH 2020*.

[253] Shammur Absar Chowdhury, Nadir Durrani, and Ahmed Ali. “What do End-to-End Speech Models Learn about Speaker, Language and Channel Information? A Layer-wise and Neuron-level Analysis”. In: *Journal of Computer Speech and Language 2021*. 2021.

[254] Tasha Nagamine, Michael L. Seltzer, and Nima Mesgarani. “Exploring How Deep Neural Networks Form Phonemic Categories”. In: *Interspeech*. 2015.

[255] Yonatan Belinkov, Ahmed Ali, and James Glass. “Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition”. In: *Interspeech 2019* (2019).

[256] Yonatan Belinkov and James Glass. “Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems”. In: *NIPS’17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017).

[257] Suwon Shon, Hao Tang, and James R. Glass. “Frame-Level Speaker Embeddings for Text-Independent Speaker Recognition and Analysis of End-to-End Model”. In: *2018 IEEE Spoken Language Technology Workshop (SLT)* (2018).

[258] Gašper Beguš and Alan Zhou. “Interpreting Intermediate Convolutional Layers of Generative CNNs Trained on Waveforms”. In: *IEEE/ACM transactions on audio, speech, and language processing*. 2022.

[259] Imen Ben-Amor et al. “Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition”. In: *Proc. INTERSPEECH 2023*. 2023, pp. 3207–3211. DOI: 10.21437/Interspeech.2023-1648.

[260] Moez Ajili et al. “Phonetic content impact on Forensic Voice Comparison”. In: *IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2016.

[261] Yun Lei et al. “A novel scheme for speaker recognition using a phonetically-aware deep neural network”. In: *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2014.

[262] Shuai Wang and Johan Rohdin. “On the Usage of Phonetic Information for Text-Independent Speaker Embedding Extraction”. In: *Interspeech*. 2019.

[263] Liang Lu et al. “The effect of language factors for robust speaker recognition”. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2009, pp. 4217–4220.

[264] Katarzyna Stąpor. “Better alternatives for stepwise discriminant analysis”. In: *Folia Oeconomica* 1.311 (2016).

[265] A. el Ouardighi, A. el Akadi, and D. Aboutajdine. “Feature Selection on Supervised Classification Using Wilks Lambda Statistic”. In: *2007 International Symposium on Computational Intelligence and Intelligent Informatics*. 2007, pp. 51–55. DOI: 10.1109/ISCIII.2007.367361.

[266] Florian Eyben, Martin Wöllmer, and Björn Schuller. “Opensmile: the munich versatile and fast open-source audio feature extractor”. In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010.

[267] Margit Antal and Gavril Toderean. “Speaker Recognition and Broad Phonetic Groups”. In: *Signal Processing, Pattern Recognition, and Applications*. 2006. URL: https://api.semanticscholar.org/CorpusID:11471553.

[268] J.P. Eatock and J.S. Mason. “A quantitative assessment of the relative speaker discriminating properties of phonemes”. In: *Proceedings of ICASSP ’94. IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. i. 1994, I/133–I/136 vol.1. DOI: 10.1109/ICASSP.1994.389337.

[269] Sandro Cumani and Salvatore Sarni. “The Distributions of Uncalibrated Speaker Verification Scores: A Generative Model for Domain Mismatch and Trial-Dependent Calibration”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023). DOI: 10.1109/TASLP.2023.3282096.

[270] Luciana Ferrer et al. “Toward Fail-Safe Speaker Recognition: Trial-Based Calibration With a Reject Option”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.1 (2019), pp. 140–153. DOI: 10.1109/TASLP.2018.2875794.

[271] Niko Brümmer. “Focal toolkit”. In: 2007. URL: google.com/site/nikobrummer/focal.

[272] Geoffrey Stewart Morrison. "Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio". In: *Australian Journal of Forensic Sciences* (2013). DOI: 10.1080/00450618.2012.733025. URL: https://doi.org/10.1080%2F00450618.2012.733025.

[273] Niko Brummer et al. "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.7 (2007), pp. 2072–2084. DOI: 10.1109/TASL.2007.902870.

[274] Niko Brümmer and George Doddington. *Likelihood-ratio calibration using prior-weighted proper scoring rules*. 2013. arXiv: 1307.7981 [stat.ML].

[275] Stéphane Pigeon, Pascal Druyts, and Patrick Verlinde. "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions". In: *Digital Signal Processing* 10.1 (2000), pp. 237–248. ISSN: 1051-2004. DOI: https://doi.org/10.1006/dspr.1999.0358. URL: https://www.sciencedirect.com/science/article/pii/S1051200499903585.

[276] Erica Gold. "Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters". PhD thesis. Jan. 2014.

[277] David van der Vloed, Finnian Kelly, and Anil Alexander. "Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA, a forensically realistic database". In: Apr. 2020. DOI: 10.21437/Odyssey.2020-57.

[278] David van der Vloed. "Data strategies in forensic automatic speaker comparison". In: *Forensic Science International* 350 (2023), p. 111790. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2023.111790.

[279] Geoffrey Stewart Morrison. "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM)". In: *Speech communication* (2011), pp. 242–256.

[280] Cuiling Zhang, Geoffrey Morrison, and Tharmarajah Thiruvaran. "Forensic voice comparison using Chinese /iau/". In: *Proceedings of the 17th International Congress of Phonetic Sciences* (Jan. 2011).

[281] Luciana Ferrer et al. "System combination using auxiliary information for speaker verification". In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2008, pp. 4853–4856. DOI: 10.1109/ICASSP.2008.4518744.

[282] Robert Tibshirani. "Regression shrinkage selection via the LASSO". In: *Journal of the Royal Statistical Society Series B* 73 (June 2011), pp. 273–282. DOI: 10.2307/41262671.

[283] Arthur E. Hoerl and Robert W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. In: *Technometrics* 42 (2000), pp. 80 –86. URL: https://api.semanticscholar.org/CorpusID:28142999.

[284] Hui Zou and Trevor J. Hastie. “Addendum: Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005). URL: https://api.semanticscholar.org/CorpusID:14134075.

[285] Ville Hautamäki et al. “Regularized logistic regression fusion for speaker verification”. In: *Proc. Interspeech 2011*. 2011, pp. 2745–2748. DOI: 10.21437/Interspeech.2011-153.

[286] Mee Young Park and Trevor Hastie. “L1-Regularization Path Algorithm for Generalized Linear Models”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69 (Aug. 2007).

[287] Mengyuan Zhang and Kai Liu. “On Regularized Sparse Logistic Regression”. In: *ArXiv* abs/2309.05925 (2023). URL: https://api.semanticscholar.org/CorpusID:261696637.

[288] Mattia Zanon et al. “Sparse Logistic Regression: Comparison of Regularization and Bayesian Implementations”. In: *Algorithms* 13 (2020).

[289] Anant Subramanian et al. *SPINE: SParse Interpretable Neural Embeddings*. 2017. arXiv: 1711.08792 [cs.CL].

[290] Marie Tahon et al. “Interprétabilité pour l’identification de locuteurs. Retour sur le projet JSALT 2023”. In: *Journée commune AFIA-TLH / AFCP*. 2023.

[291] Daniel Lee and H. Seung. “Learning the Parts of Objects by Non-Negative Matrix Factorization”. In: *Nature* 401 (Nov. 1999), pp. 788–91. DOI: 10.1038/44565.

[292] Brian Murphy, Partha Pratim Talukdar, and Tom Michael Mitchell. “Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding”. In: *International Conference on Computational Linguistics*. 2012. URL: https://api.semanticscholar.org/CorpusID:8348149.

[293] Alona Fyshe et al. “Interpretable Semantic Vectors from a Joint Model of Brain- and Text- Based Meaning”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Kristina Toutanova and Hua Wu. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 489–499. DOI: 10.3115/v1/P14-1046. URL: https://aclanthology.org/P14-1046.

[294] Shy Shoham, Daniel O’Connor, and Ronen Segev. “How silent is the brain: Is there a "dark matter" problem in neuroscience?” In: *Journal of comparative physiology. A, Neuroethology, sensory, neural, and behavioral physiology* 192 (Sept. 2006), pp. 777–84. DOI: 10.1007/s00359-006-0117-6.

[295] Edward Kim et al. *The Interpretable Dictionary in Sparse Coding*. 2020. arXiv: 2011.11805 [cs.LG].

[296] Hoagy Cunningham et al. *Sparse Autoencoders Find Highly Interpretable Features in Language Models*. 2023. arXiv: 2309.08600 [cs.LG].

[297] Oriol Vinyals and Li Deng. "Are sparse representations rich enough for acoustic modeling?" In: *Proc. Interspeech 2012*. 2012, pp. 2570–2573. DOI: 10.21437/Interspeech.2012-8.

[298] Minxue Xia and Hao Zhu. *Interpretable Neural Embeddings with Sparse Self-Representation*. 2023. arXiv: 2306.14135 [cs.CL].

[299] Diego Garcia-Olano et al. *Intermediate Entity-based Sparse Interpretable Representation Learning*. 2022. arXiv: 2212.01641 [cs.CL].

[300] Andrew Ng. "Sparse autoencoder". In: *CS294A Lecture notes* 72 (2011), 1–19.

[301] Miguel Á. Carreira-Perpiñán and Ramin Raziperchikolaei. "Hashing with binary autoencoders". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 557–566. DOI: 10.1109/CVPR.2015.7298654.

[302] Francisco Mena and Ricardo Ñanculef. "A Binary Variational Autoencoder for Hashing". In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2019, 131–141. ISBN: 978-3-030-33903-6. DOI: 10.1007/978-3-030-33904-3_12. URL: https://doi.org/10.1007/978-3-030-33904-3_12.

[303] Viacheslav Osaulenko. "Binary autoencoder with random binary weights". In: *ArXiv* abs/2004.14717 (2020). URL: https://api.semanticscholar.org/CorpusID:216869181.

[304] Mohamed Amine Hmani, Dijana Petrovska-Delacrétaz, and Bernadette Dorizzi. "Locality preserving binary face representations using auto-encoders". In: *IET Biometrics* 11.5 (2022), pp. 445–458. DOI: https://doi.org/10.1049/bme2.12096. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/bme2.12096.

[305] Haoyu Wang et al. "Task-Agnostic Structured Pruning of Speech Representation Models". In: *Proc. INTERSPEECH 2023*. 2023, pp. 231–235. DOI: 10.21437/Interspeech.2023-1442.

[306] Dinghan Shen et al. *Learning Compressed Sentence Representations for On-Device Text Processing*. 2019. arXiv: 1906.08340 [cs.CL].

[307] Yi-Hsuan Tsai et al. *Learning Binary Residual Representations for Domain-specific Video Streaming*. 2017. arXiv: 1712.05087 [cs.CV].

[308] Ruslan Baynazarov and Irina Piontkovskaya. "Binary Autoencoder for Text Modeling". In: *Communications in Computer and Information Science* (2019). URL: https://api.semanticscholar.org/CorpusID:209070068.

[309] Yoshua Bengio, Nicholas Leonard, and Aaron Courville. “Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation”. In: *arXiv preprint arXiv:1308.3432 (2013)*. 2013.

[310] Penghang Yin et al. “UNDERSTANDING STRAIGHT-THROUGH ESTIMATOR IN TRAINING ACTIVATION QUANTIZED NEURAL NETS”. In: *ICLR 2019*. 2019.

[311] Shuchang Zhou et al. “DOREFA-NET: TRAINING LOW BITWIDTH CONVOLUTIONAL NEURAL NETWORKS WITH LOW BITWIDTH GRADIENTS”. In: *Computer Science*. 2018.

[312] Jure Zbontar et al. *Barlow Twins: Self-Supervised Learning via Redundancy Reduction*. 2021. arXiv: 2103.03230 [cs.CV].

[313] Carlos Busso et al. “IEMOCAP: Interactive emotional dyadic motion capture database”. In: *Language Resources and Evaluation* 42 (Dec. 2008), pp. 335–359. DOI: 10.1007/s10579-008-9076-6.

[314] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[315] Patrick Meyer and Gianluca Bontempi. “On the Use of Variable Complementarity for Feature Selection in Cancer Classification”. In: Apr. 2006, pp. 91–102. ISBN: 978-3-540-33237-4.