Université d'Avignon
Laboratoire Informatique d'Avignon

Ph.D dissertation

Deep modeling based on voice attributes for explainable speaker recognition

Application in the domain of forensics

Imen Ben-Amor

Supervised by:

Pr. Jean-François Bonastre

Supported and funded by:

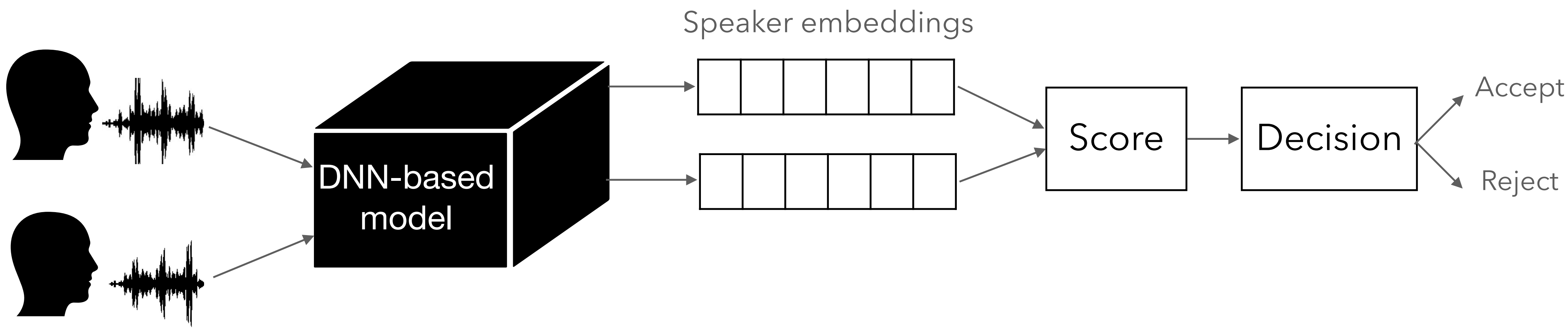
L AVIGNON
CHAIRE PARTENARIALE EN INTELLIGENCE ARTIFICIELLE

25/04/2024

Plan

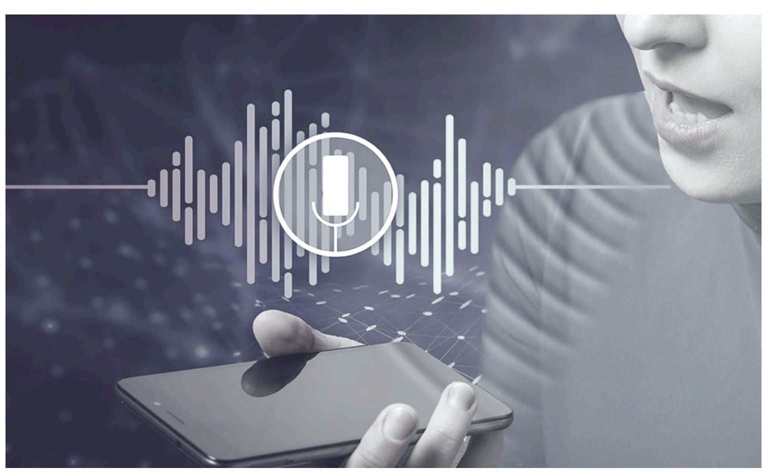
- I Context & motivations
- II BA-LR: proposed methodology
- III Application on forensic data
- IV Modelling improvements
- V Conclusion and perspectives

Automatic Speaker Recognition (ASpR)



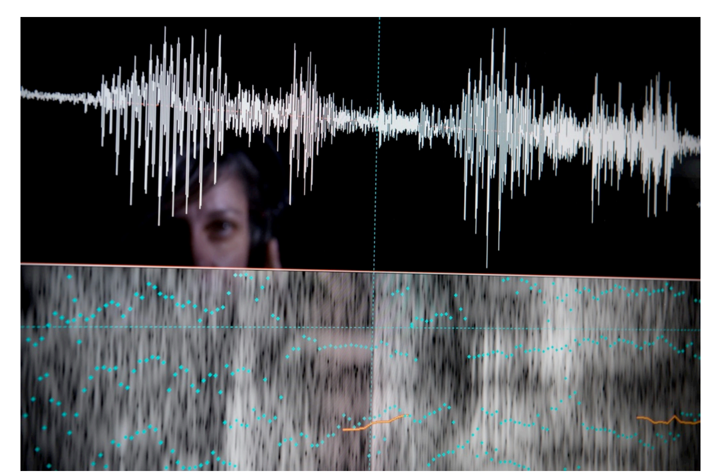
Applications

Smart assistant

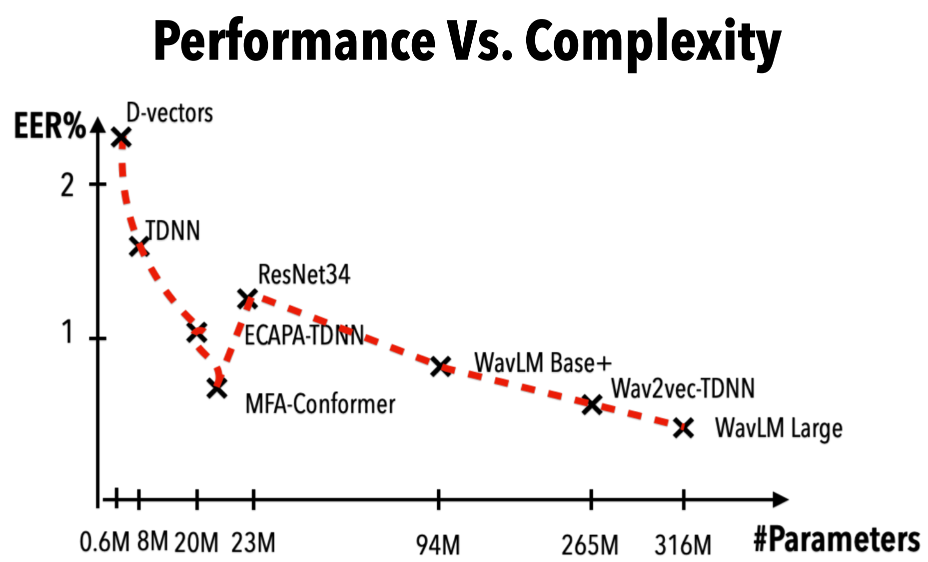


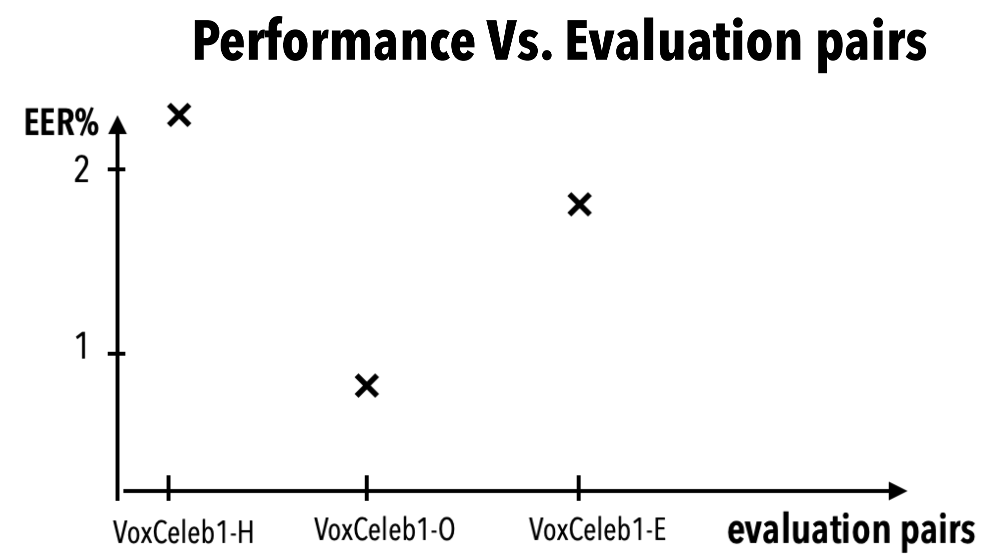
Biometric authentication

Forensics



State of the art ASpR





- Higher performance.
  - ➔ More complex architectures and higher number of parameters.
  - ➔ Sacrify the interpretability of the information flow.
- The variability of train data + The non representative choice of evaluation pairs + Speech quality.
  - ➔ Unpredictable output.
  - ➔ A risk of discrimination bias [Khoury2013, Hutiri2022].

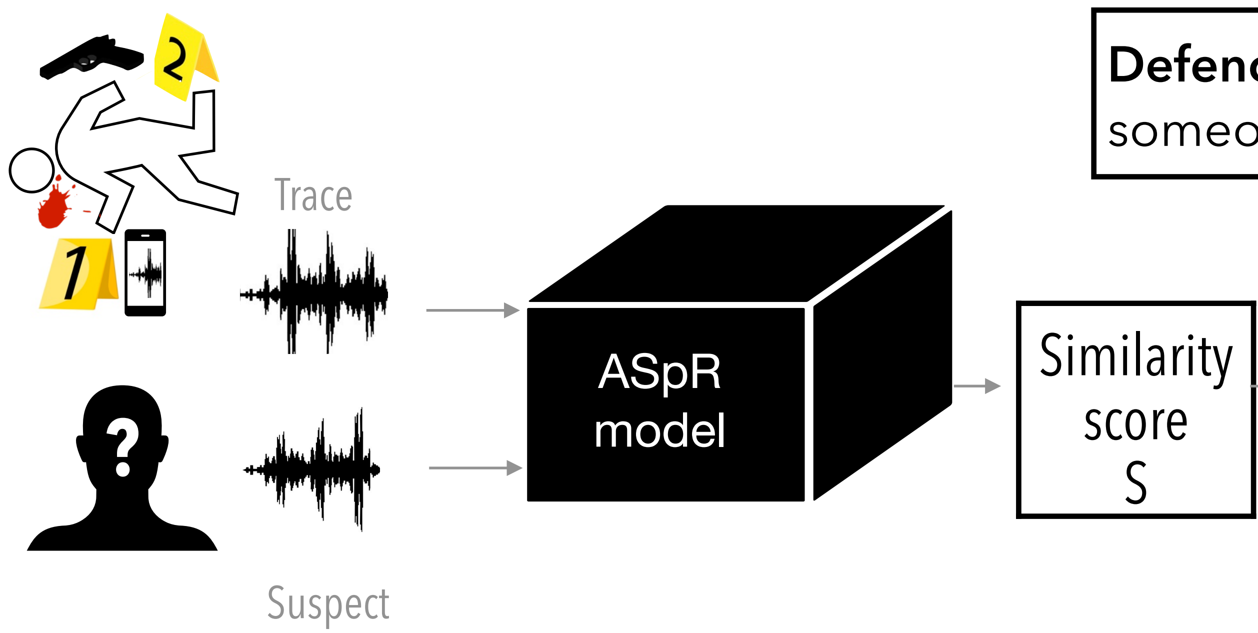
☞ **In this thesis, we aim to address the opacity of ASpR models and provide well informed output applied in forensic context.**

Forensic automatic speaker recognition

Centrality of likelihood ratio

Prosecution hypothesis (H_p): The trace sample belongs to the suspect.

Defence hypothesis (H_d): The trace sample belongs to someone other than the suspect.



$$LR = \frac{P(S_{Trace,Suspect} \mid H_p)}{P(S_{Trace,Suspect} \mid H_d)}$$

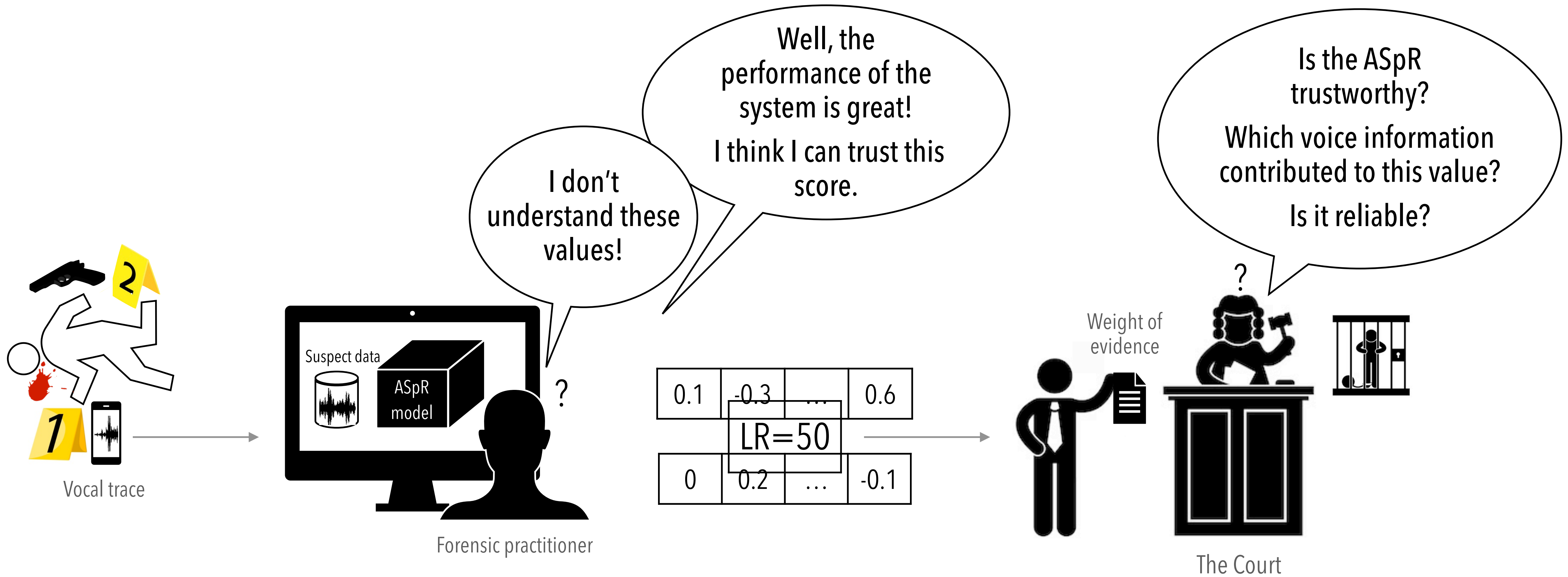
Bayes paradigm

$$\underbrace{\frac{P(H_p \mid E)}{P(H_d \mid E)}}_{\text{Posterior odds}} = \underbrace{\frac{P(E \mid H_p)}{P(E \mid H_d)}}_{\text{Likelihood ratio}} * \underbrace{\frac{P(H_p)}{P(H_d)}}_{\text{Prior odds}}$$

Value of evidence

The lack of interpretability

Forensic context



☞ **System performance alone is not enough to trust a DNN-based model.**

☞ **The interpretability and the transparency of the output produced by the system is a <u>MUST</u>** [Deeks2019, Solanke2022, Kirat2023].

Research questions

❖ This thesis aims to propose an interpretable and explainable ASpR approach.

- **RQ1**: Can we make the embedding space interpretable?
- **RQ2**: Which voice information influences the final score in ASpR task? What is its contribution? Is it reliable?
- **RQ3**: What is the nature of this encoded information?

Our inspiration

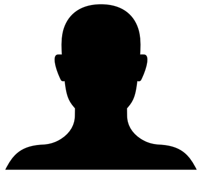
Simplified forensic DNA identification

Identification process

Trace genotype

Locus 5 — 5

Locus 12 — 12



Suspect profile

Locus 5 — 5

Locus 12 — 12

Many locus: Loci

Typicality: presence frequency in a given population.

$$\text{Partial-LR}_{locus_5} = \frac{\text{P(Trace matches the suspect)}}{\text{P(Trace matches someone else in population)}} = \frac{\text{Similarity}_{Gene_5}}{\text{Typicality}_{Gene_5}}$$

$$LR = \prod \text{partial-LR}_{locus_i}$$

Uncertainty in a locus [Gill2008, Shestak2021]

Drop-out

Trace genotype

Locus 5 — 5

Locus 12 — 12

Drop-in

Trace genotype

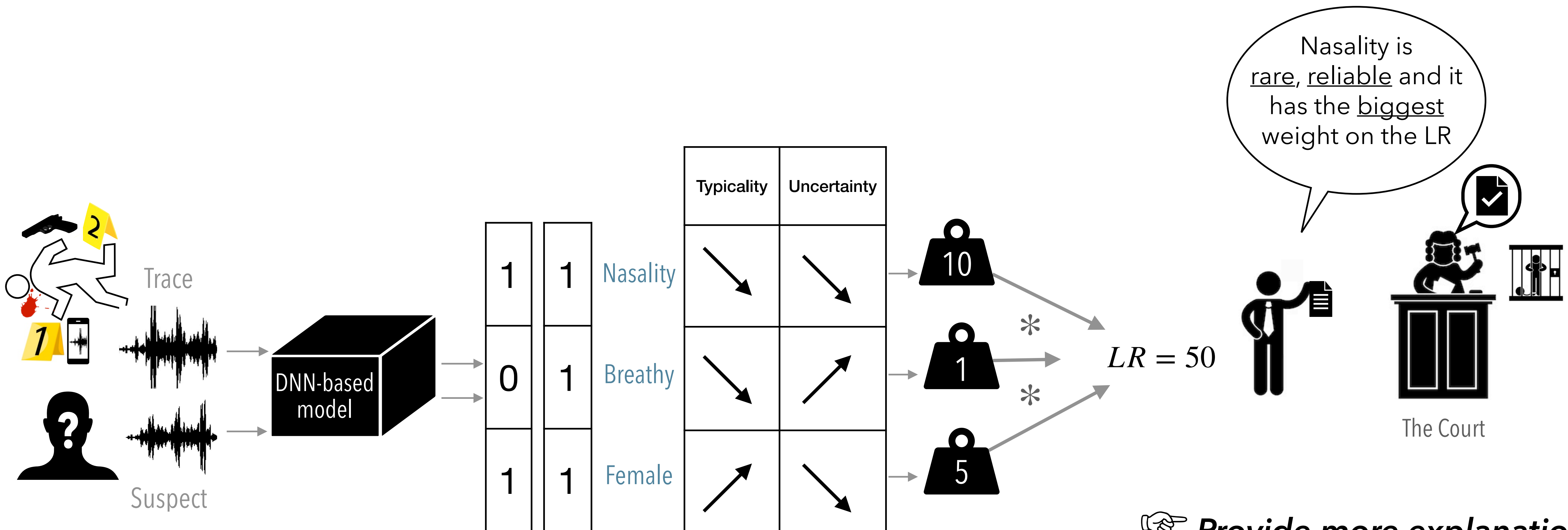
Locus 5 — 5

Locus n — N

→ These are considered in partial-LRs [Gill2012]

Proposed ideal solution

What if?

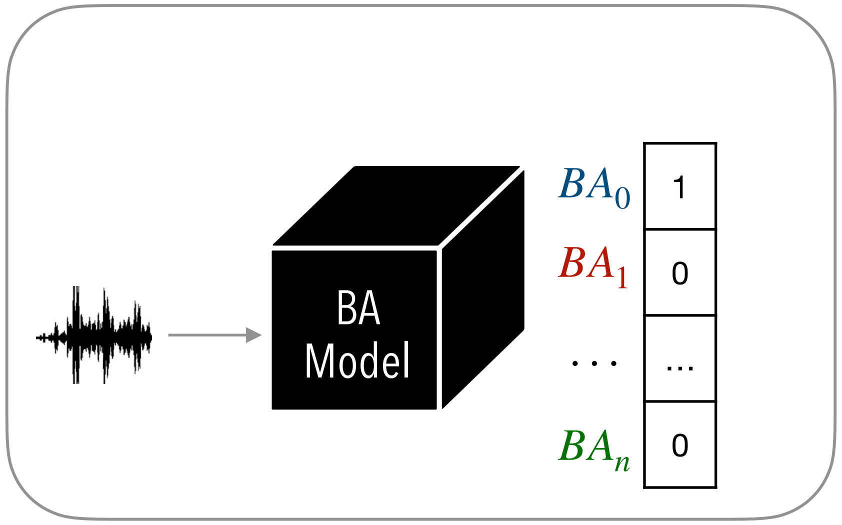


☞ ***Provide more explanations.***

☞ ***Allow a better handle of the value of evidence.***

BA-LR three-step methodology

Step 1



Step 2

	BA_0	BA_1	...	BA_n
Typicality	0.1	0.5	...	0.9
Uncertainty	0.7	0.5	...	0.1

$$LR = LR^{BA_0} * LR^{BA_1} * \ldots * LR^{BA_n}$$

Step 3

	BA_0	BA_1	...	BA_n
Phonemes	Vowels	Nasals	...	Stops
Phonetics	F0	Formants	...	Energy

1. **Binary and attribute-based modelling**: Represent a speech sample by a binary vector, where each dimension represents the presence or absence of an assumed attribute.

2. **Interpretable and explainable scoring**: Decompose the LR as the product of attribute-LRs, each associated to an attribute.

3. **Attribute explainability**: Describe the nature of attributes in terms of phonetic and phonemic information.

STEP 1: Binary and attribute-based speaker embeddings

Binary attribute-based modelling

- Related work on binary speaker embeddings
  - Preserve privacy and enhance security of speaker information [Boufounos2011]
  - Reduce both time and computational costs [Li2016]
  - Model speaker specific discriminant information [Bonastre2011]

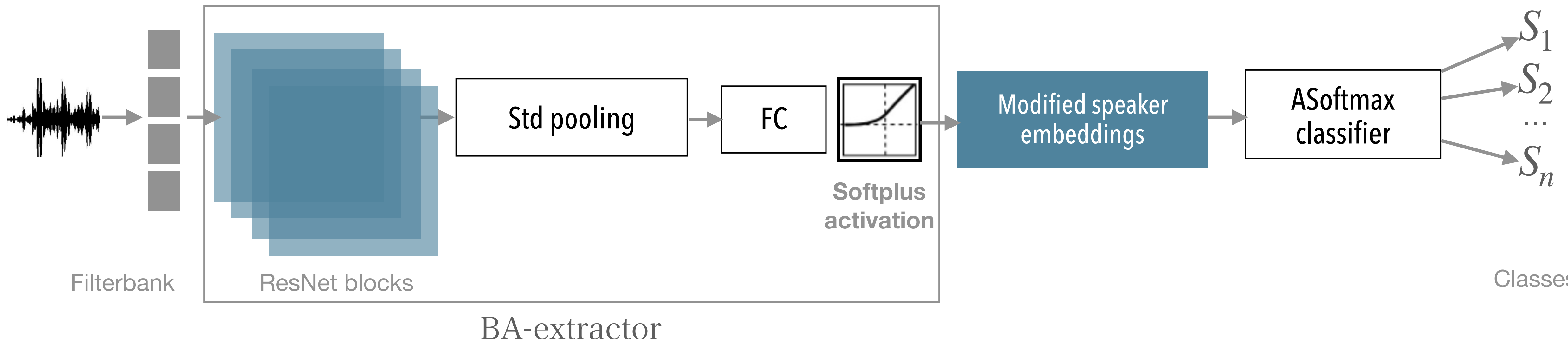
☞ **Our goal is to model binary and attribute-based speaker embeddings, assuming:**

- A speech sample is represented by the presence (1) or absence (0) of <u>predefined</u> set of attributes.
- Attributes are <u>shared</u> between groups of speakers.
- Attributes are assumed to be <u>independent</u>.

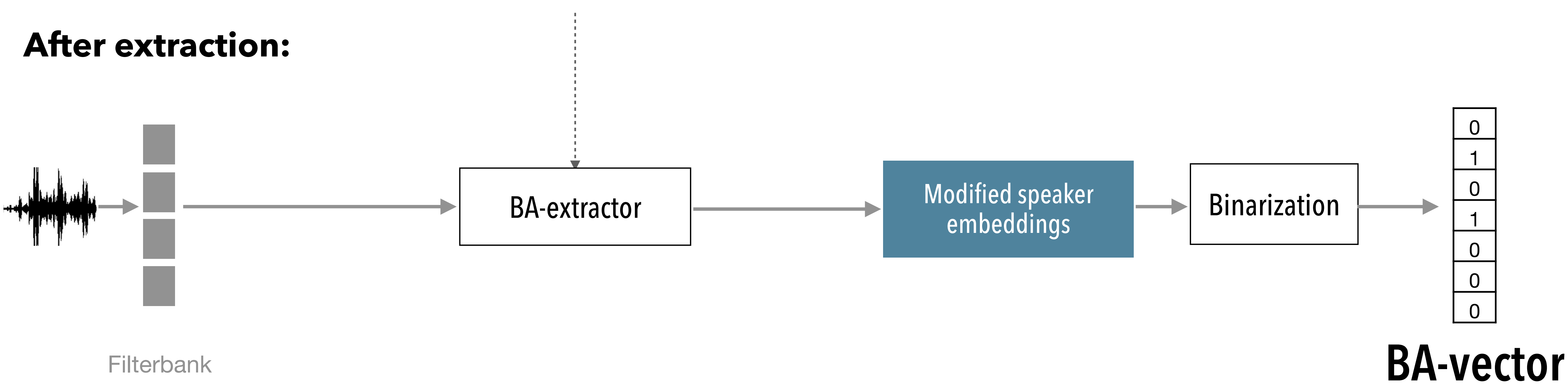
BA-extractor model

- The proposed model is based on a modified ResNet extractor [Zeinali2019].

During training:



After extraction:



ASpR performance

- **Datasets**

Datasets description

	VoxCeleb2	*VoxCeleb1*
	Train	Evaluation
# of speakers	5,994	1,251
# of extracts	1,021,175	153,516
# of test pairs		56,295*2

The number of pairs is balanced between target and non-target

- **Evaluation using Cosine similarity**

ASpR Performance in terms of EER on VoxCeleb1

	X-vectors	*BA-vectors*
# of dimensions	256 floats (8192 bits)	205 bits
EER	1.37 %	3.42 %

EER: Intersection point between FAR and FRR

☞ Good ASpR performance.

☞ A ~2% of absolute increase in EER compared to x-vectors.

☞ A dimensionality reduction of x-vectors by ~40 times.

Key takeaways

- Represent speech samples by binary vectors, modelled by voice attributes <u>shared</u> among speakers.
- Adds a thresholding function to orient the representations towards binarization.

✓ A good trade-off between binarization and ASpR performance.

✗ ResNet architecture is not the most accurate.

✗ The post-extraction binarization is not ideal.

STEP 2: BA-LR

Binary-Attribute-based Likelihood Ratio estimation

Existing LR estimation methods

- **Score-based methods** [Bolck2015, Leegwater2017]: $LR = \dfrac{f(S_{X,Y} \mid H_p)}{f(S_{X,Y} \mid H_d)}$

✓ Widely used and easily implemented.

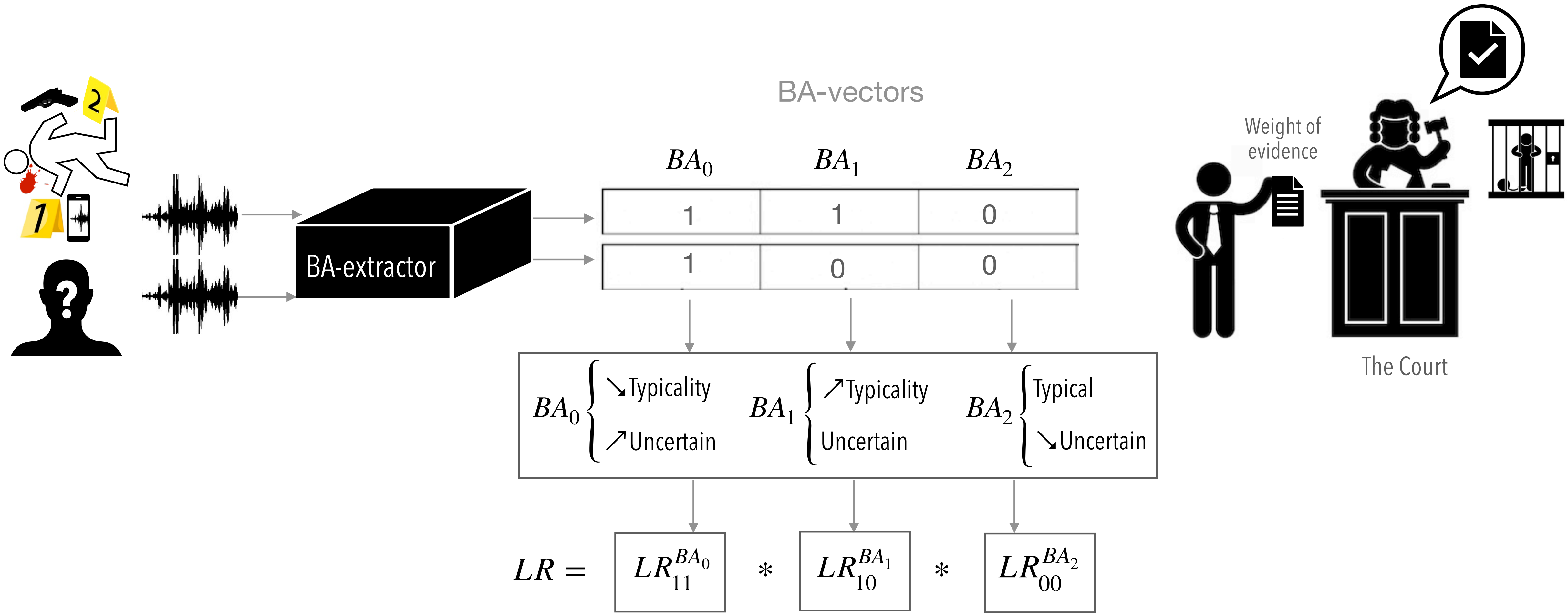
✗ Reduce the multivariate feature vectors to a compact single similarity score.

- **Feature-based methods** [Franco-Pedroso2016]: $LR = \dfrac{f(x, y \mid H_p)}{f(x, y \mid H_d)}$

✓ Consider the similarity as well the typicality of feature vectors under comparison.

✗ Consider the entire distribution but not each feature contribution to the LR.

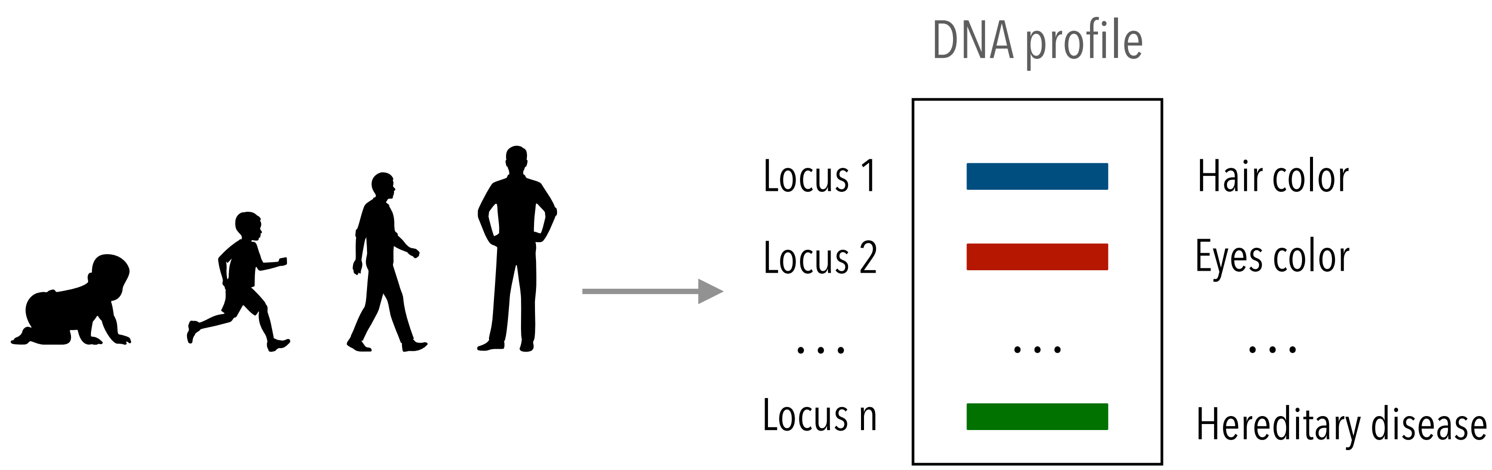
Interpretable BA-LR scoring

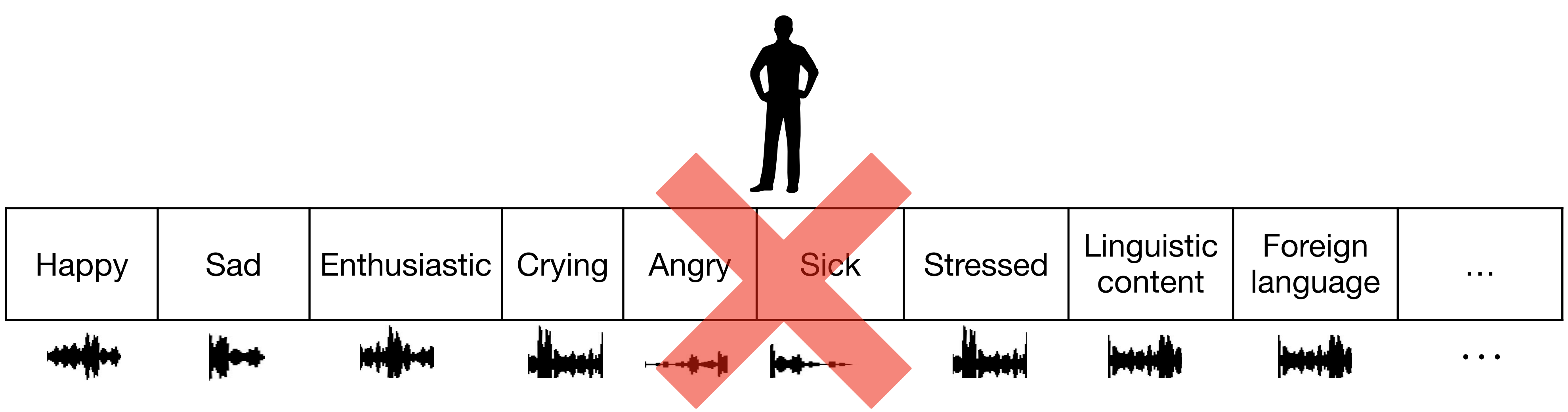


- RQ1: How to estimate the behavior of each attribute?
- RQ2: How to estimate an interpretable LR per attribute?
- RQ3: Is BA-LR applicable in an ASpR task?
- RQ4: Which explanations does it offer to the final LR?

Estimation of behavioral parameters

The "elusive" speaker profile



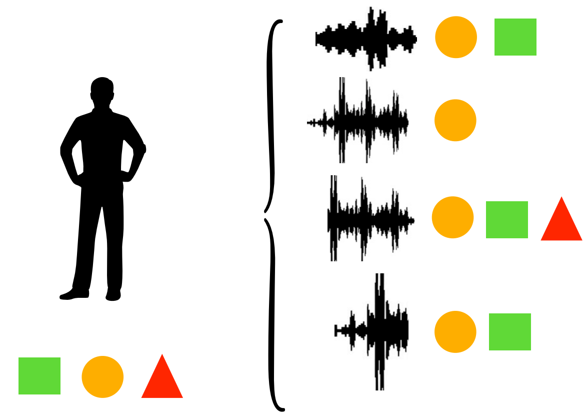


The speaker profile is a myth

Estimation of behavioral parameters

The "elusive" speaker profile

- The attribute is present in the profile if it is present at least once in the available set of speaker utterances.



Speaker Profile

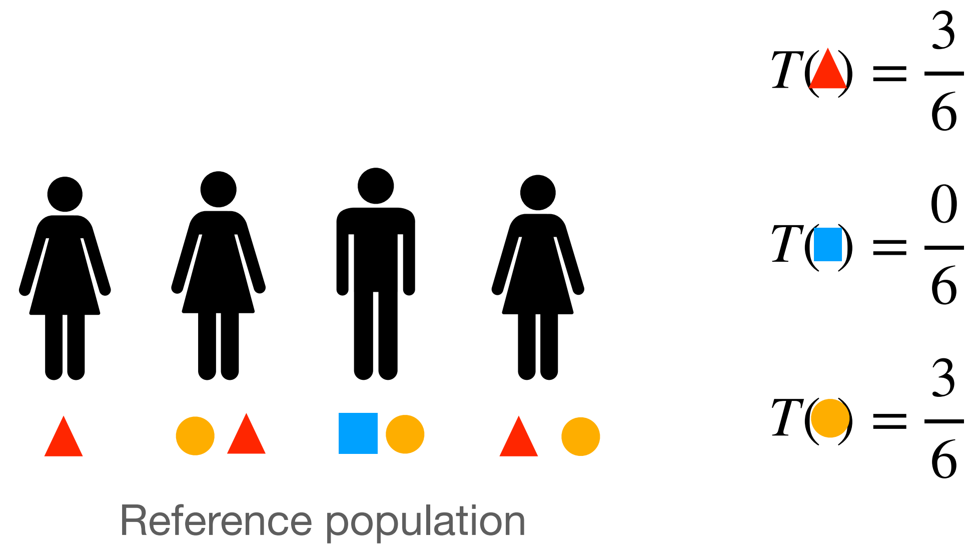
Estimation of behavioral parameters

Typicality

The frequency of speaker pairs in the reference population sharing the attribute in their profiles.

$$T(BA_i) = \frac{\sum^{N_c} P_{S1} \cap P_{S2} = \{BA_i = 1\}}{N_c}$$

P_{S_j} is the speaker profile



The reference population is the set of speakers from the training data of the DNN model [Drygajlo et al].

Estimation of behavioral parameters

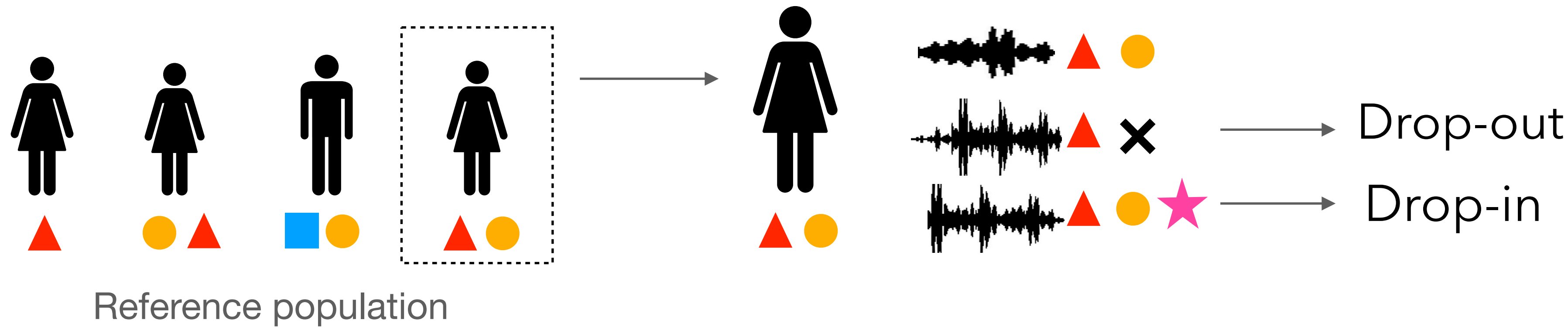
Uncertainty: Drop-out & Drop-in

Drop-out - disappearance of attribute: occurs due to a false negative detection or due to a non presence of the attribute.

$$Dout_i^S = \frac{\sum_{U \in S}^{N_S} (U(BA_i = 0) \,|\, P_S(BA_i) = 1)}{N_S} \qquad Dout_i = \frac{\sum_j^N Dout_i^{S_j}}{N}$$

Drop-in - appearance of foreign attribute: occurs due to a false positive detection of the attribute.

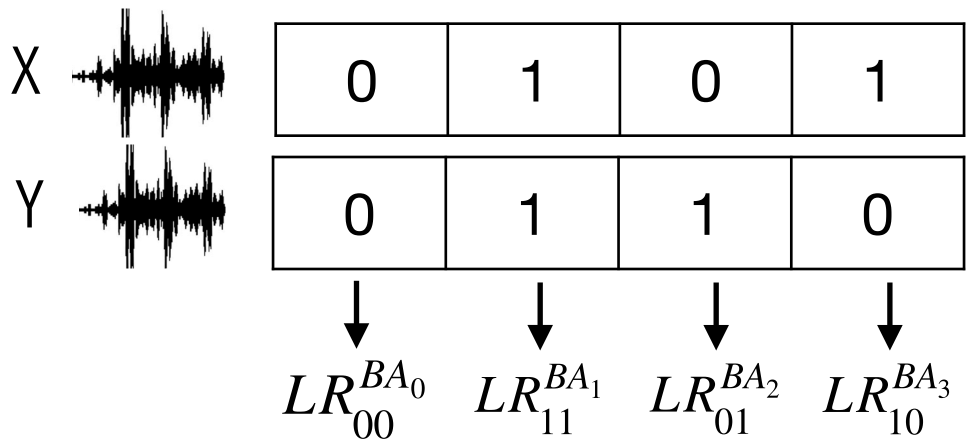
$$\text{Dropin}_i = Din * T_i$$



Din: Estimate speech noise

Interpretable attribute-LR estimation

Speech-adapted BA-LR



$$\mathrm{LR}^{BA_i}_{X_i,Y_i} = \frac{P(X_i, Y_i \,|\, H_p)}{P(X_i, Y_i \,|\, H_d)}$$

Assumptions:

- Drop-in and drop-out could occur in X and Y.
- Both phenomena are independent.

T_i: Typicality | $\overline{Din}$: No drop-in |$\overline{Dout}$: No drop-out

❖Case $X_i = 1$, $Y_i = 1$

11→11 00→11

Same speaker

Under H_p

$$\mathrm{LR}^{BA_i}_{X_i,Y_i} = \frac{1 + (\mathrm{Din} \cdot T_i)^2}{T_i \cdot (\overline{\mathrm{Dout}_i}^{\,2} + (\mathrm{Din} \cdot T_i)^2 + 2 \cdot \mathrm{Din} \cdot T_i \cdot \overline{\mathrm{Dout}_i})}$$

Under H_d

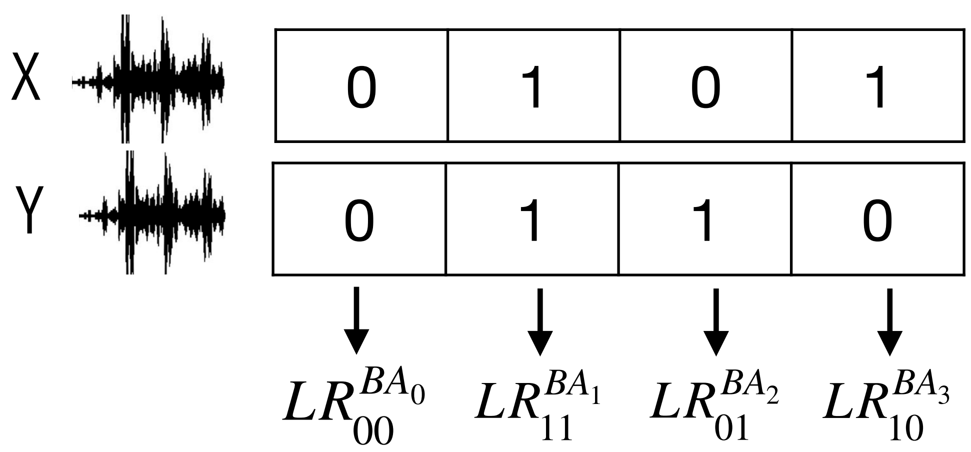
11→11 00→11 01→11 10→11

Different speakers

Imen Ben Amor and Jean-François Bonastre, "BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison," In: **IWBF2022.**

Interpretable attribute-LR estimation

Speech-adapted BA-LR



$$\mathrm{LR}_{X_i,Y_i}^{BA_i} = \frac{P(X_i, Y_i \mid H_p)}{P(X_i, Y_i \mid H_d)}$$

$$\begin{cases} \dfrac{1+\mathrm{Dout}_i^2}{T_i\cdot\left(2\cdot\mathrm{Dout}_i\cdot\overline{\mathrm{Din}}+\mathrm{Dout}_i^2+\overline{\mathrm{Din}}^2\right)} & \mathrm{if}(\mathrm{BA}_i^Y=0,\mathrm{BA}_i^X=0)\\ \dfrac{1+\left(\mathrm{Din}\cdot T_i\right)^2}{T_i\cdot\left(2\cdot\mathrm{Din}\cdot T_i\cdot\overline{\mathrm{Dout}_i}+\left(\mathrm{Din}\cdot T_i\right)^2+\overline{\mathrm{Dout}_i}^2\right)} & \mathrm{if}(\mathrm{BA}_i^Y=1,\mathrm{BA}_i^X=1)\\ \dfrac{\overline{\mathrm{Din}}\cdot\mathrm{Din}\cdot T_i+\mathrm{Dout}_i\cdot\overline{\mathrm{Dout}_i}}{T_i\cdot\left(\overline{\mathrm{Din}}\cdot\mathrm{Din}\cdot T_i+\mathrm{Dout}_i\cdot\overline{\mathrm{Dout}_i}+1+\mathrm{Din}\cdot T_i\cdot\mathrm{Dout}_i\right)} & \mathrm{if}(\mathrm{BA}_i^Y=0,\mathrm{BA}_i^X=1)\\ \dfrac{\overline{\mathrm{Din}}\cdot\mathrm{Din}\cdot T_i+\mathrm{Dout}_i\cdot\overline{\mathrm{Dout}_i}}{T_i\cdot\left(\overline{\mathrm{Din}}\cdot\mathrm{Din}\cdot T_i+\mathrm{Dout}_i\cdot\overline{\mathrm{Dout}_i}+1+\mathrm{Din}\cdot T_i\cdot\mathrm{Dout}_i\right)} & \mathrm{if}(\mathrm{BA}_i^Y=1,\mathrm{BA}_i^X=0) \end{cases}$$

Assumptions:

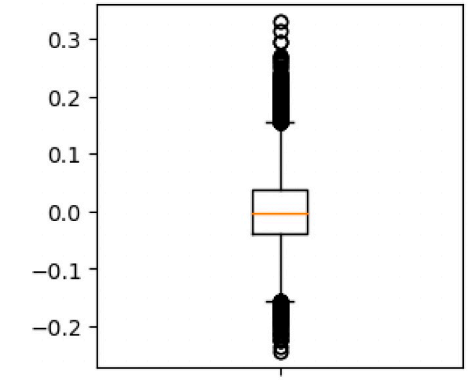
- Drop-in and drop-out could occur in X and Y.
- Both phenomena are independent.

T_i: Typicality | $\overline{Din}$: No drop-in | $\overline{Dout}$: No drop-out

Imen Ben Amor and Jean-François Bonastre, "BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison," In: **IWBF2022.**

ASpR performance

- Small correlation between attributes in BA-vectors.



Correlation values between BAs

ASpR performance evaluated on three datasets in terms of EER and Cllr

	X-vectors		***BA-vectors***	
	Cosine		Speech-adapted BA-LR	
	EER	$Cllr_{min/act}$	EER	$Cllr_{min/act}$
VoxCeleb1	1.37 %	0.06 / 0.82	3.5 %	0.13 / 0.48
SITW (Wild conditions)	1.4 %	0.06 / 0.82	4 %	0.14 / 0.49
VOiCES (Challenging environment)	3.96 %	0.15 / 0.87	5.12 %	0.19 / 0.89

Cllr is the cost associated with the log LR decision threshold
EER: Equal error rate. (Lower is better)

☞ A good ASpR performance and generalisation ability using BA-LR scoring.

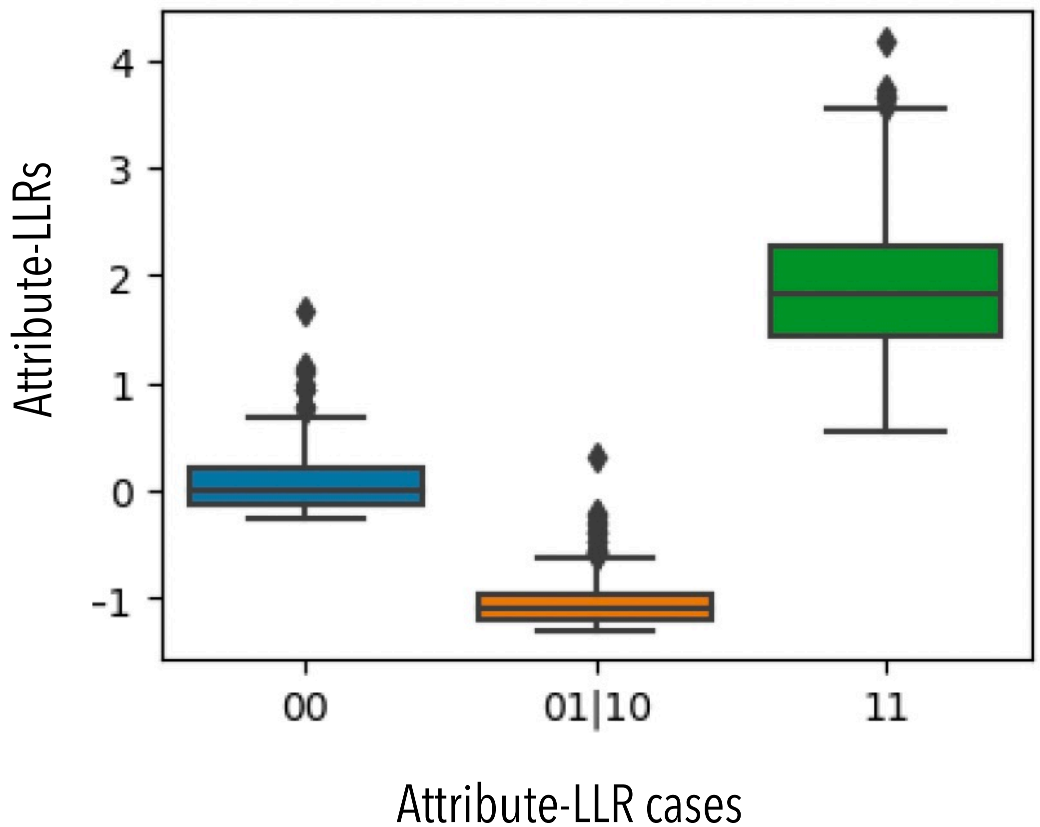
☞ An average increase of 1.96% in EER compared to x-vectors.

☞ Poorly calibrated LRs.

Interpretability of attribute LLRs

$$LLR = Log(LR) = \sum \text{attribute-LLR}_i$$

- The case 00 gives **<u>very small</u>** attribute-LLRs ➔ Negligible impact on the LLR.
- In the case 01 or 10, the attribute-LLRs are all **<u>negative</u>** ➔ A conflict that decreases the LLR.
- The case 11 gives **<u>positive and high</u>** attribute-LLRs ➔ Adds an important weight to the LLR.



Explainability of the LLRs

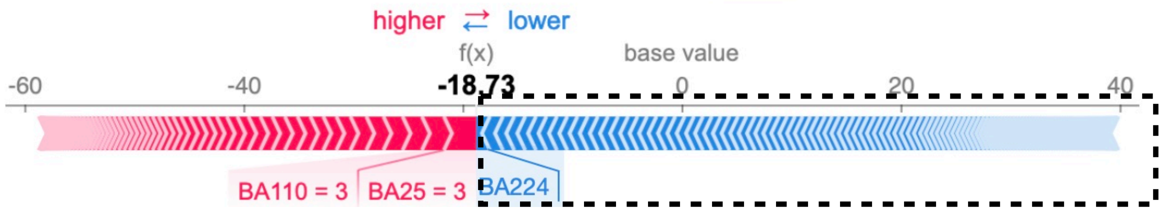
Shapley-like explanations

$$LLR = Log(LR) = \sum \text{attribute-LLR}_i$$

- Contribution of attribute= attribute-LLR
- For target, there is more attributes pushing the final LLR towards positive direction.
- For non-target, there is more attributes pushing the final LLR towards negative direction.
- The most contributing attributes are characterized by a low typicality and an acceptable drop-out.

Target example

Non-Target example



	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		

https://github.com/shap/shap

Key takeaways

- Establish an interpretable and explainable computation of the LR in an ASpR task.
- A transparent BA-LR scoring based on a <u>simplified estimation</u> of behavioral parameters, allowing a better handle of the value of evidence.

✓ Good ASpR performance and generalisation abilities.

✓ BA-LR provides explanations about the contribution of each attribute to the final LLR.

✗ The notion of speaker profile is misleading.

✗ The estimation of behavioral parameters is limited.

✗ ASpR performance might be not sufficient enough for some applications.

STEP 3: Attribute explainability

Existing explainability methods

- Use probing classifiers and available labels to investigate speaker information within the embeddings [Wang2017, Raj2019].
- An analysis of the phonemic information along neural network layers [Nagamine2015].

Our prerequisites:

- Attributes are derived from a bottom-up extractor.
- No information is available about the nature of these attributes.

A solution that ensures:

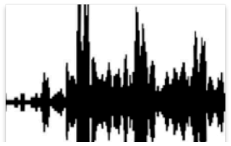
- No additional labelling or annotation of data.
- Cover all cases from the train data.
- Automatic discovery and description of attributes.

Proposed explainability method

The three-world method

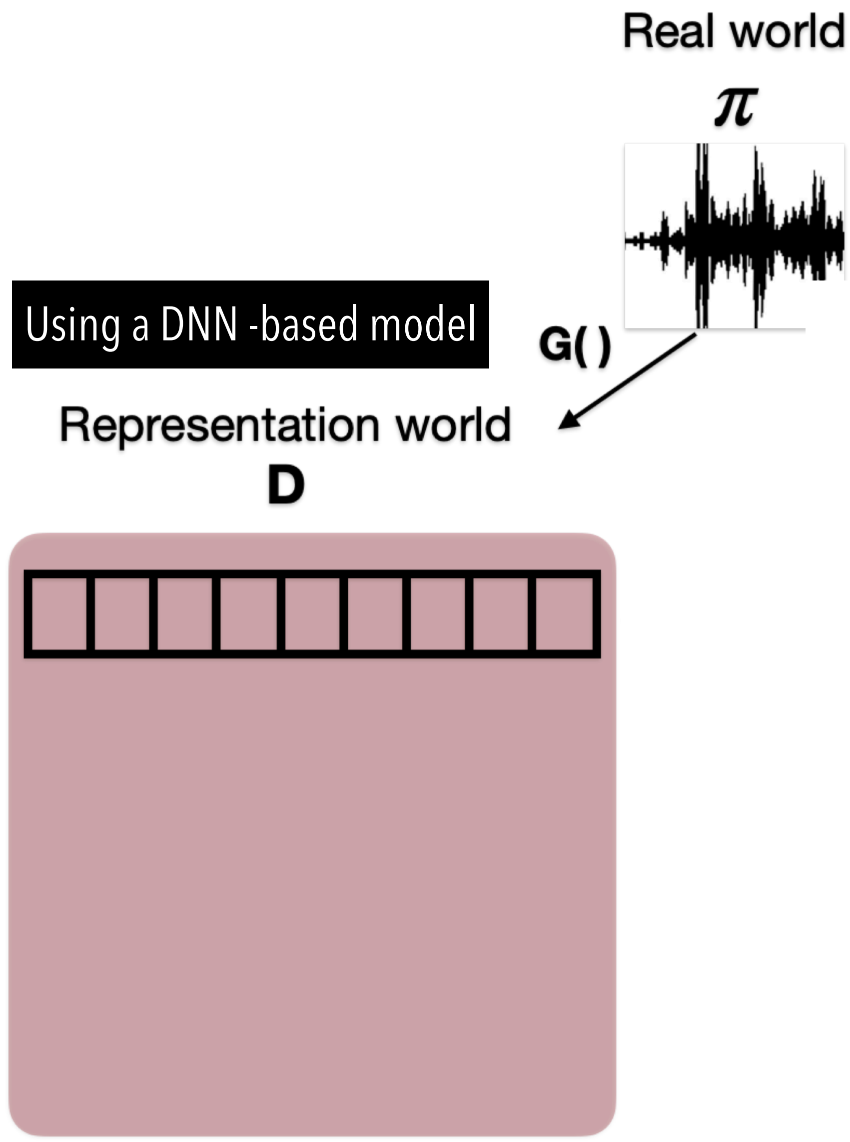
Real world

π



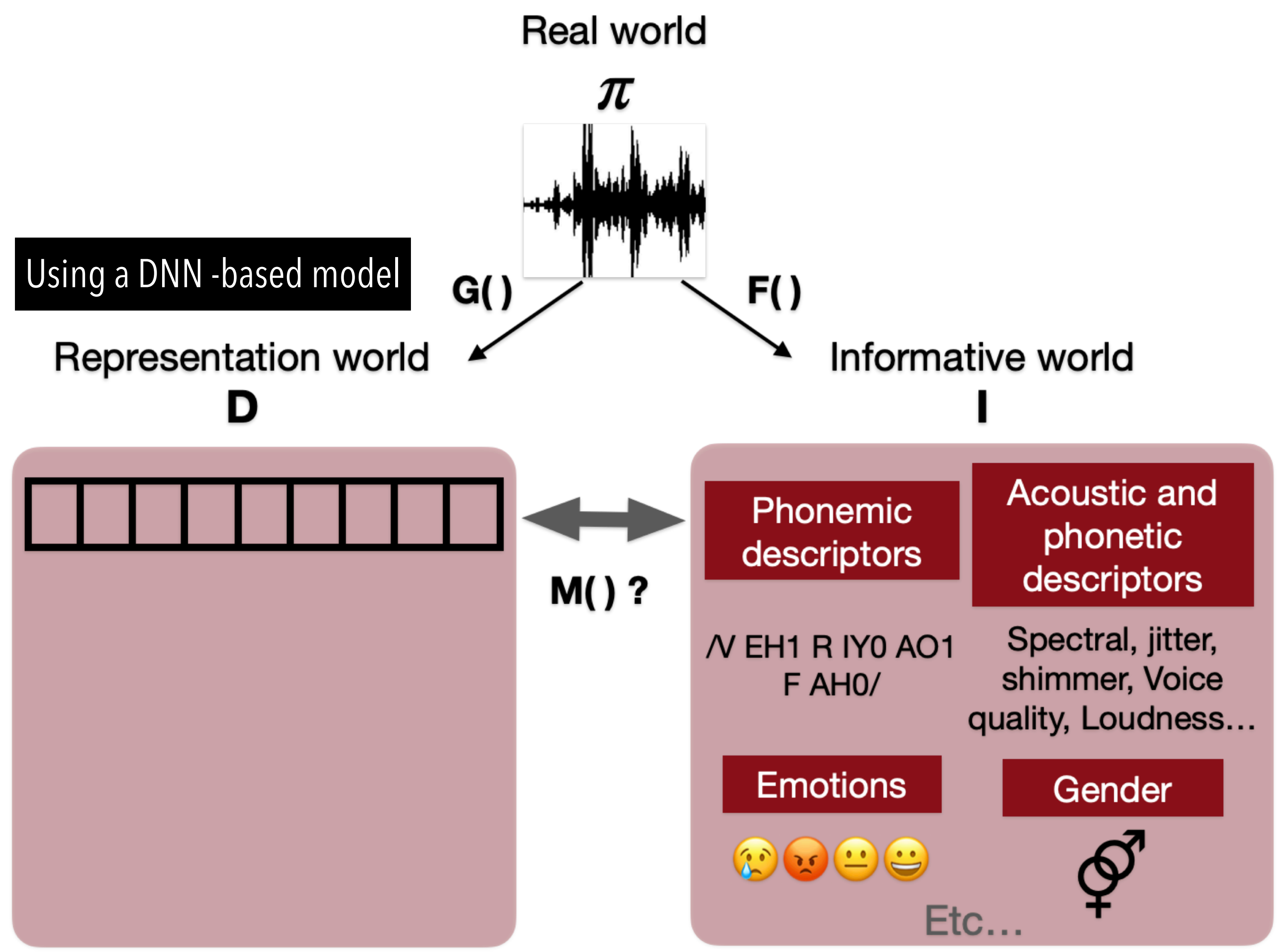
Proposed explainability method

The three-world method



Proposed explainability method

The three-world method



How to determine an automatic mapping M() between D and I?

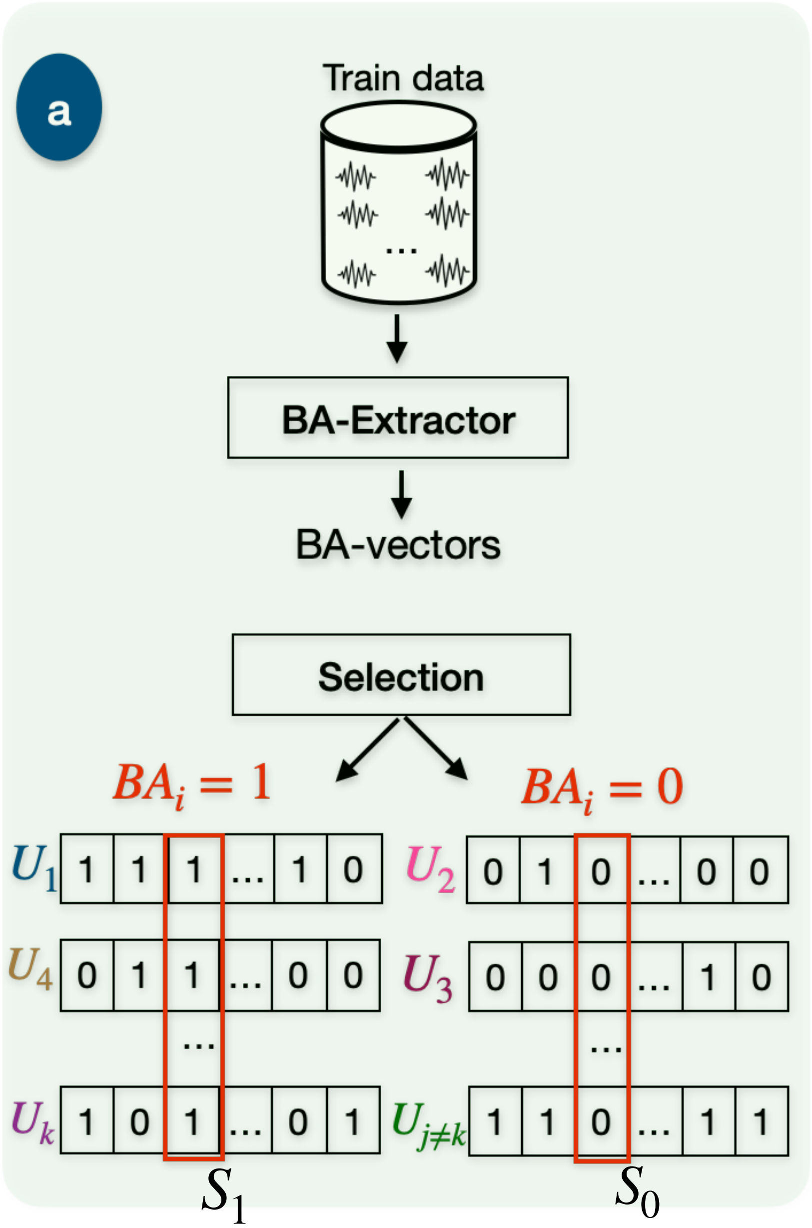
Utterance-level mapping

Methodology

Assumption: *If variables in the I world are able to differentiate between the 0/1 of an attribute in the D world, then these variables are good descriptors of the attribute.*

Thanks to binarization, for each attribute:

a. Select speech samples and group them in two sets: S_0 where attribute is 0 and S_1 where attribute is 1.



Imen Ben-Amor et.al, "Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition",In: **Interspeech 2023**

Utterance-level mapping

Methodology

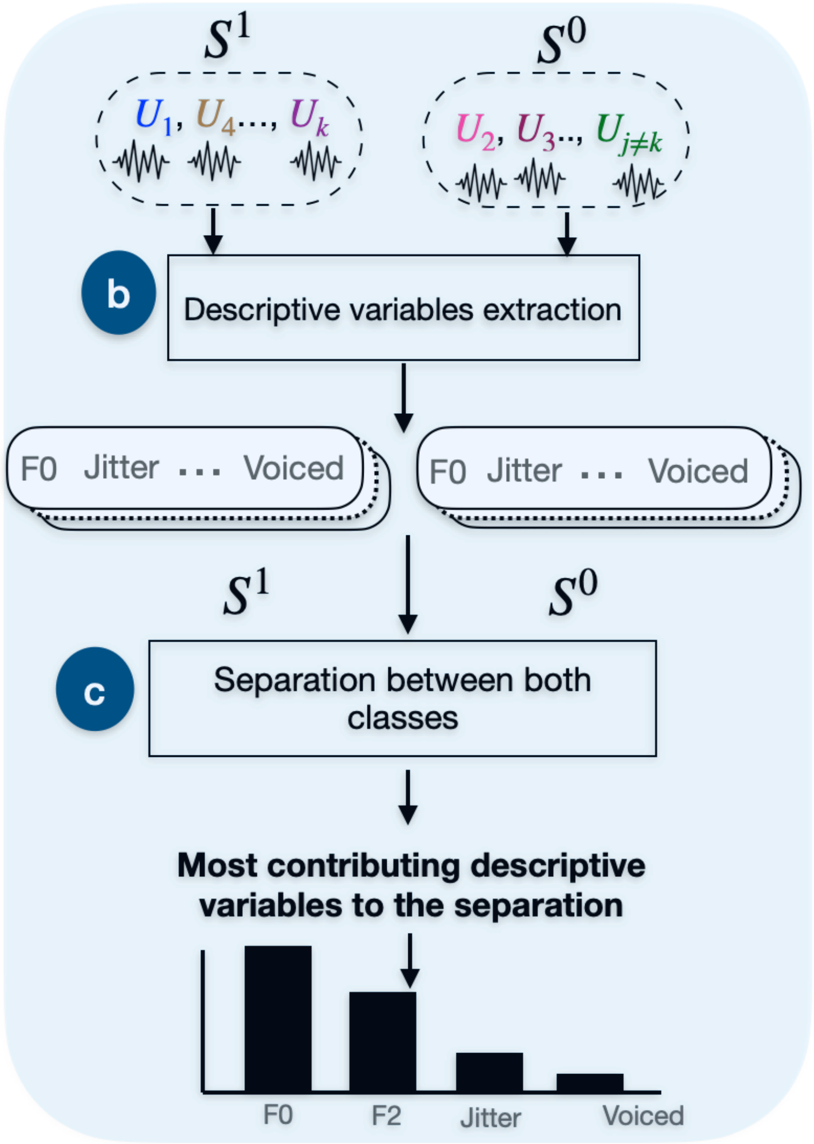
<u>Assumption:</u> *If variables in the I world are able to differentiate between the 0/1 of an attribute in the D world, then these variables are good descriptors of the attribute.*

<u>Thanks to binarization</u>, for each attribute:

a. Select speech samples and group them in two sets: S_0 where attribute is 0 and S_1 where attribute is 1.

b. Extract descriptive variables from the speech samples of both sets.

c. Separate between S_0 and S_1 via a mapping function and choose the best descriptive variables for this separation.



Utterance-level mapping

Mapping functions

1. **A surrogate model: an inherently interpretable classifier**

   - Decision Tree classifier: takes phonetic descriptive variables and predicts the presence (class=1) or absence (class=0) of the attribute in the D world.

   - TreeShap: Selects the most contributing variables to the separation between the two classes.

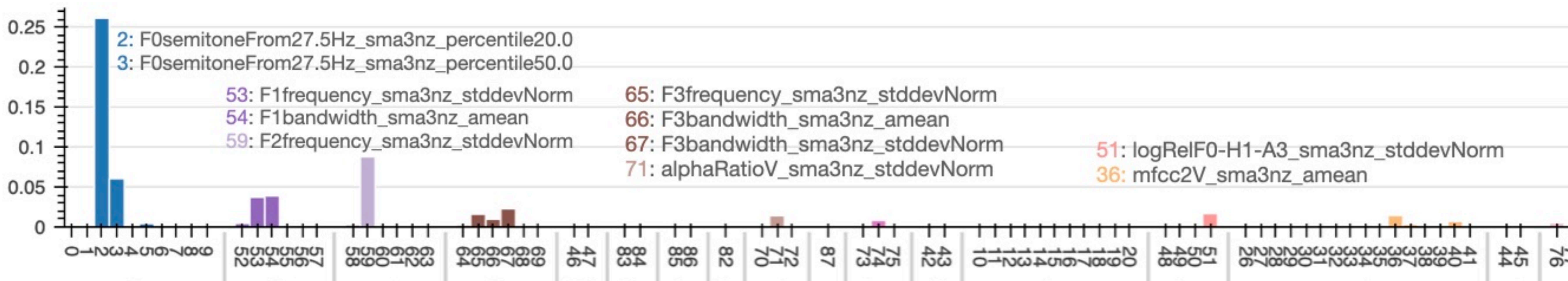
2. **Stepwise linear discriminant analysis (SLDA)**: selects a subset of the most discriminant variables to separate the two classes of the attribute.

https://github.com/shap/shap

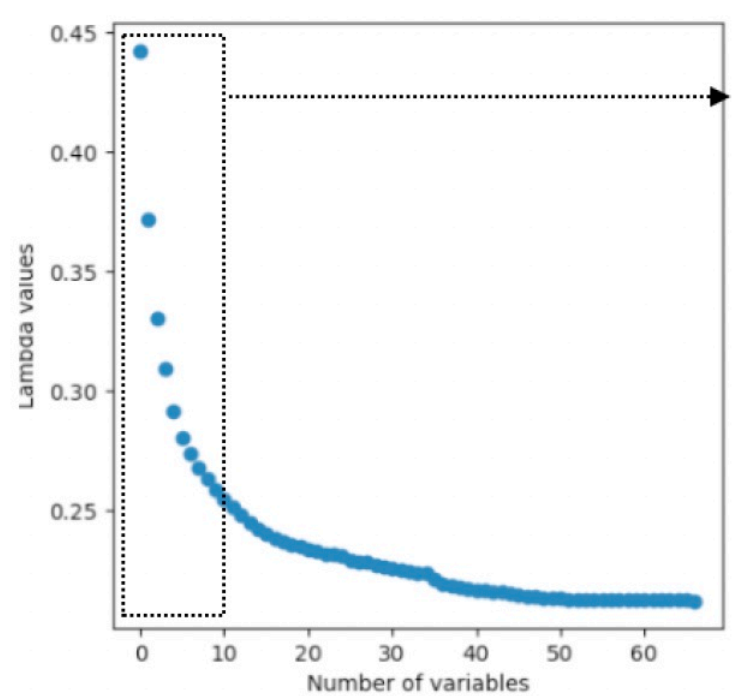
Phonetic description

Example attribute BA9

- Using Decision Tree+ TreeShap



- Using SLDA

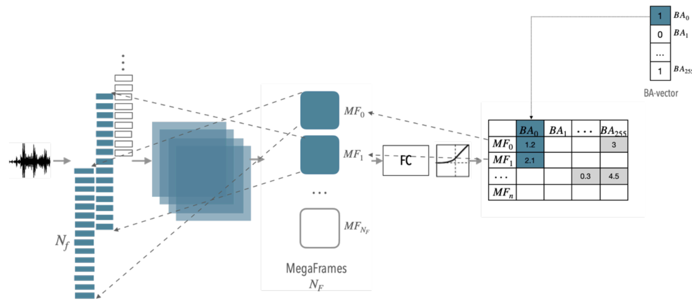


First 10 discriminant variables

- F0semitoneFrom27.5Hz_sma3nz_percentile50
- F2frequency_sma3nz_stddevNorm
- F1bandwidth_sma3nz_amean
- logRelF0-H1-A3_sma3nz_amean
- F1frequency_sma3nz_stddevNorm
- F3bandwidth_sma3nz_stddevNorm
- shimmerLocaldB_sma3nz_amean
- loudness_sma3_percentile50
- slopeV0-500_sma3nz_amean
- F2bandwidth_sma3nz_stddevNorm

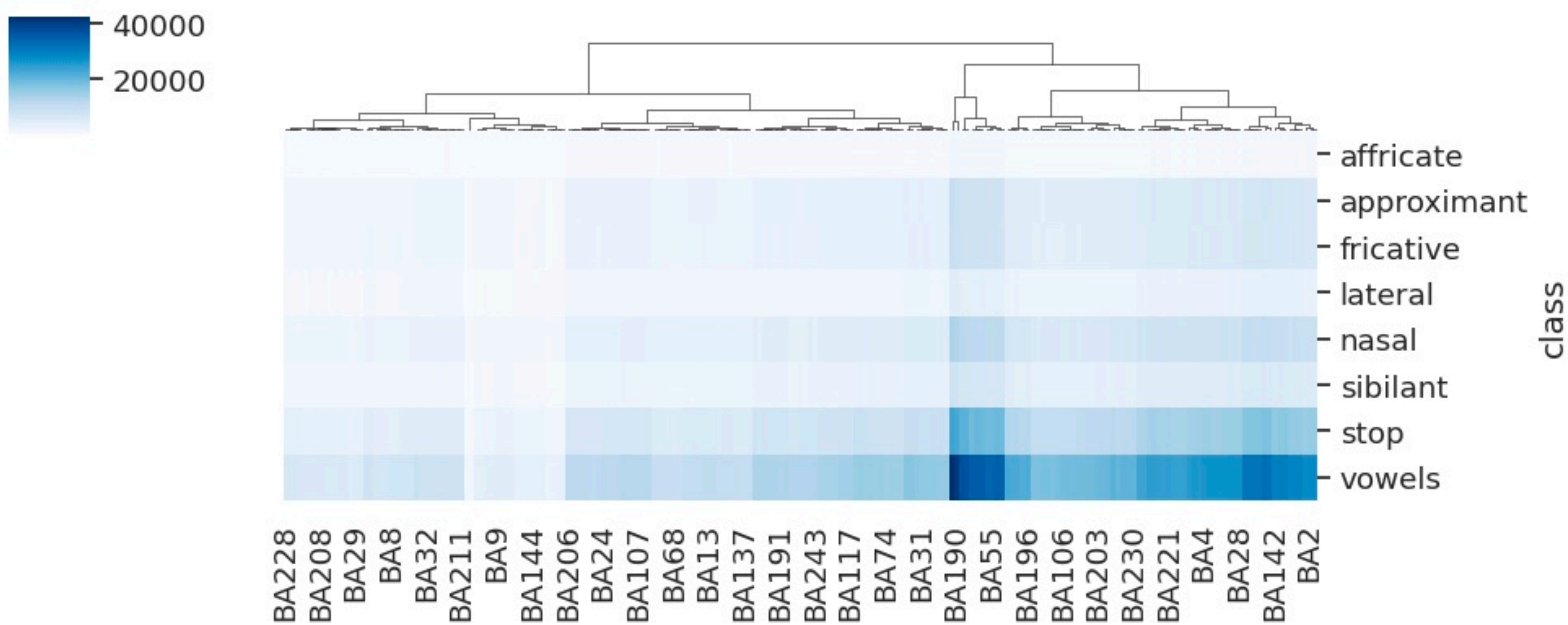
Frame-level: phonemic description

Mapping: attributes ↔ phonemes



☞ Vowels are mostly selected, followed by the Stops and the Nasals.

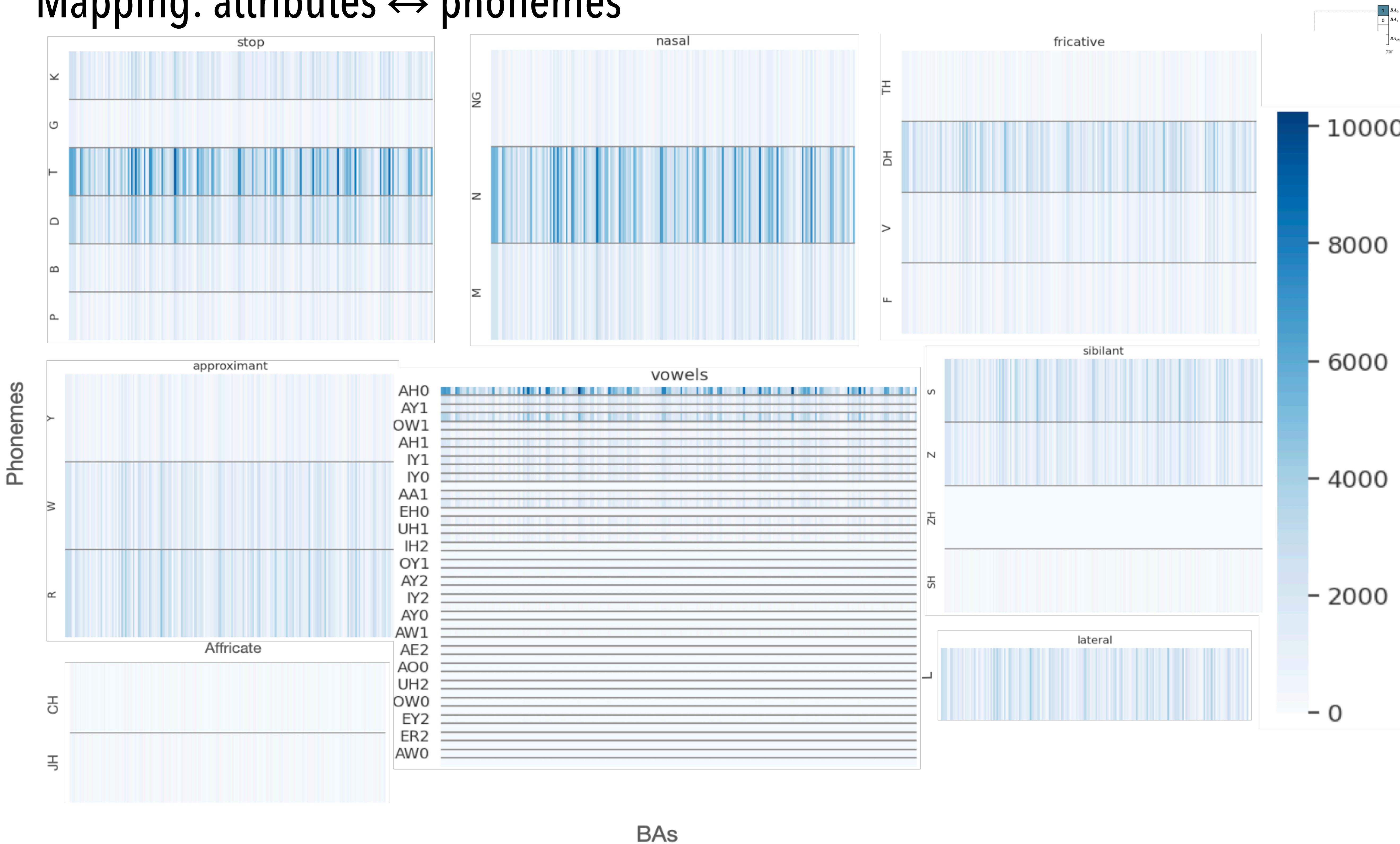
☞ In [Shon2018, Antal2006] vowels and nasals are shown important for speaker discrimination.



Occurence of each class of phonemes, clustered per BAs

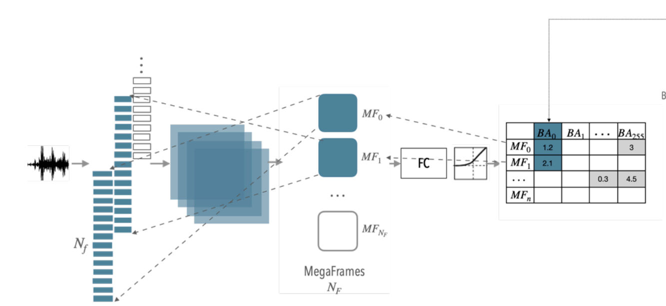
Frame-level: phonemic description

Mapping: attributes ↔ phonemes



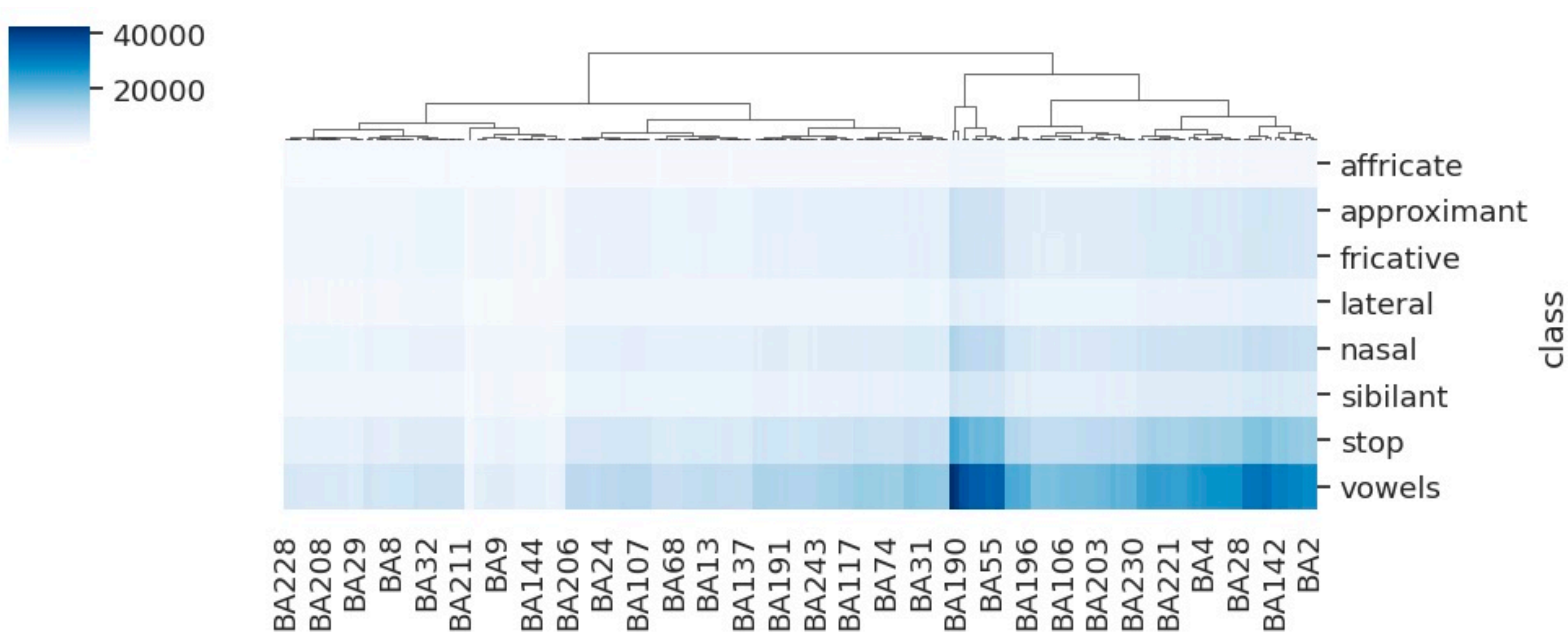
Frame-level: phonemic description

Mapping: attributes ↔ phonemes



☞ Vowels are mostly selected, followed by the Stops and the Nasals.

☞ In [Shon2018, Antal2006] vowels and nasals are shown important for speaker discrimination.



Occurence of each class of phonemes, clustered per BAs

Key takeaways

- Explain and describe the nature of information encoded within attributes.
- An automatic mapping through two levels between attributes and phonetic and phonemic descriptions.

✓ Attributes encode distinct phonetic and phonemic information.

✓ Descriptions provide insightful explanations.

✓ A useful tool helping phoneticians to discover new combinations of descriptors.

✗ A lack of a higher-level interpretation for non-experts in phonetics.

Application on forensically realistic data

Forensically realistic data: NFI-FRIDA

Data description

During my visit to the NFI in September 2023.

Netherlands Forensic Institute
Ministry of Justice and Security

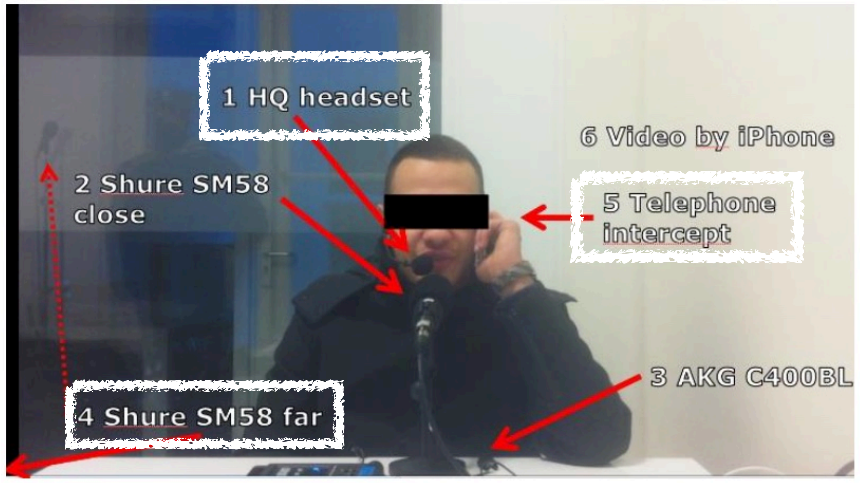
- A Dutch speech database recorded by 302 male participants via forensically significant devices.

- **Devices**: we focus on 3 devices

  - **Device d1**: Headset microphone with high quality.

  - **Device d4**: Low quality police interview recordings.

  - **Device d5**: intercepted telephone recordings.



- **Sessions:** Inside-silent/noisy, outside-calm/busy street

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**

The need for calibration

In such a forensic context:

- Mismatch in domain, conditions and population between train and evaluation data.

We remind also that:

- The BA-extractor is trained on VoxCeleb2, a predominantly English dataset.
- The behavioral parameters of BA-LR are also calculated on VoxCeleb2.

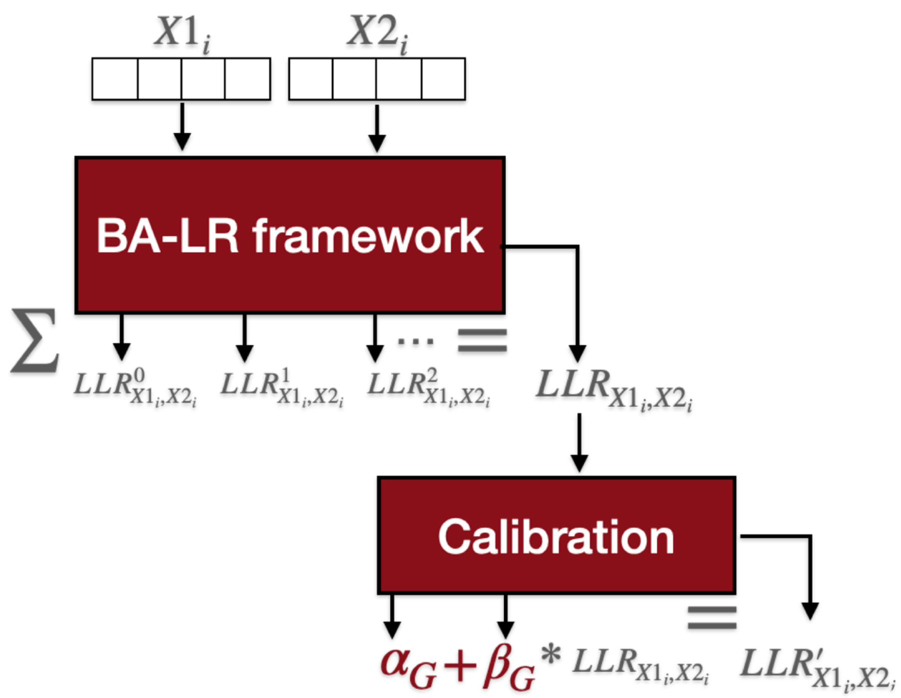
☞ **This mismatch may lead to poorly calibrated LLRs.**

☞ **A calibration step is needed!**

Calibration and fusion methods

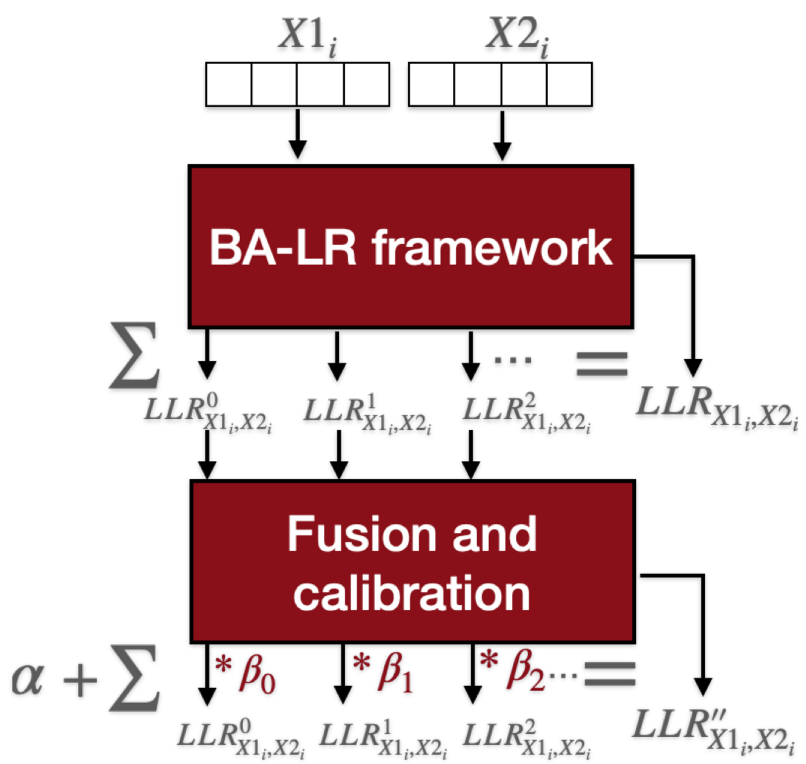
Global calibration of final LLRs

- Univariate Logistic Regression
- Shift and scale the final LLRs
- Improve calibration

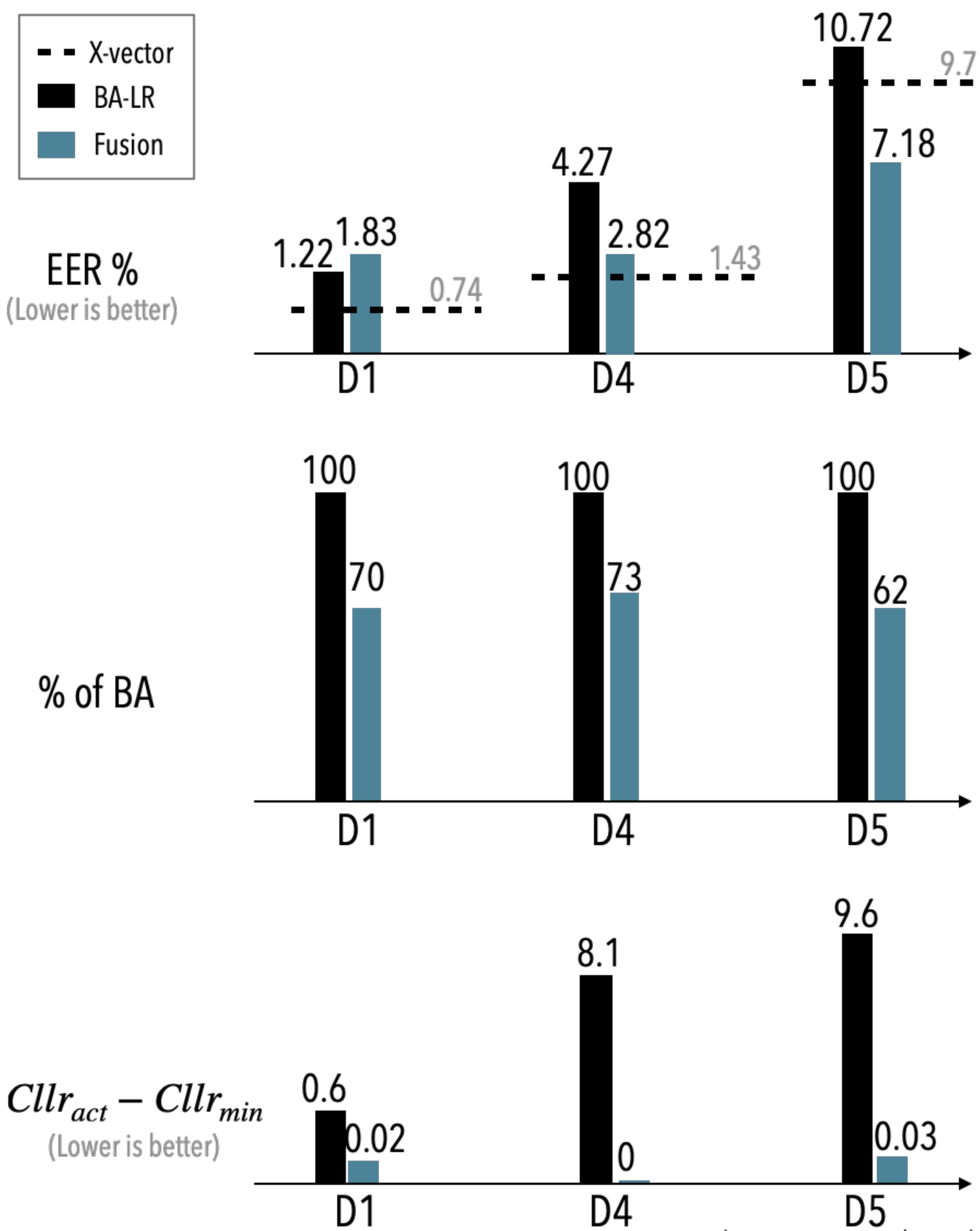


Weighted fusion of attribute-LLRs

- Multivariate Logistic Regression
- Sparse regularization
- Select only relevant attribute-LLRs
- Alleviate the independence assumption between attributes.



ASpR performance and calibration



- Divide each data device into dev and test.
- Train the calibration on dev and evaluate ASpR on Test.

☞ Generalisation ability of BA-LR scoring.

☞ The fusion **improved** the ASpR performance using BA-LR scoring.

☞ A slight increase in EER for d1.

☞ The fusion selects **only** ~**70%** of 205 attributes.

☞ Both methods **effectively calibrated** the initially miscalibrated LLRs.

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**

Key takeaways

- Address the LLRs miscalibration using BA-LR scoring on forensically realistic dataset.
- A Logistic Regression model is applied on LLRs for calibration + for an optimal fusion of attribute-LLRs.

✓ Generalisation ability of BA-LR on Dutch data.

✓ This fusion improved both calibration and ASpR performance.

✗ Further research is still needed for a forensic real world deployment.

Modelling improvements:
Attribute-based binary auto-encoder

Limitations of the BA-extractor

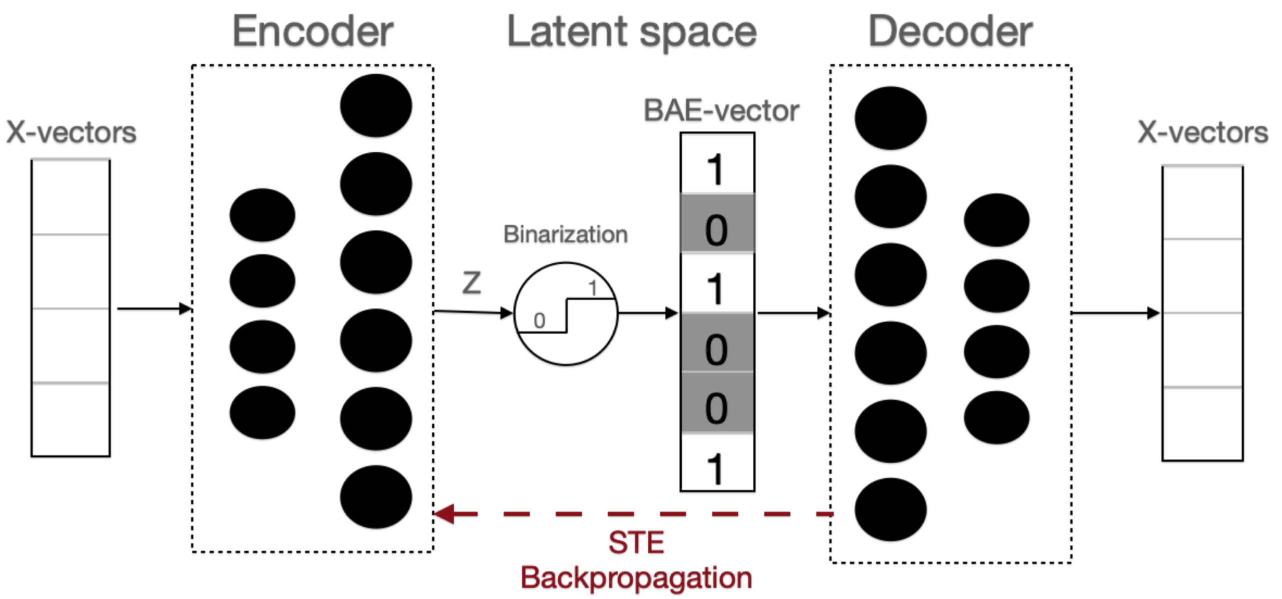
- The binarization aspect is not integrated into the modelling.
- The objective of shared attribute is not directly considered.
- The ASpR performance declines compared to x-vectors.

☞ **Explore a new direction based on auto-encoder architecture.**

BAE: Attribute-based binary auto-encoder

Architecture

- **Input**: x-vectors of 256 dimensions.
- **Latent space**: BAE-vector of 512 dimensions.
- **Forward**: z is binarized converting negative values to 0 and positive to 1.
- **Backward**: the gradient back-propagate using StraightThrough Estimator [Bengio2013].



BAE: Attribute-based binary auto-encoder

Proposed attribute-oriented loss

- Encourage the shared attribute behavior in the binary vectors.
- By controlling the presence frequency of attributes among speakers.
- This refers to the concept of typicality where an attribute may be <u>rare</u>, <u>moderately present</u> or <u>typical</u> among speakers.

Desired typicality

BA_0	BA_1	BA_2	BA_3
0.1	0.3	0.5	0.9

During training

	BA_0	BA_1	BA_2	BA_3
Profile speaker 1	✓	✓	✓	
Profile speaker 2		✓	✓	✓
Profile speaker 3			✓	
Profile speaker 4	✓	✓	✓	✓

$$L_S = \sum (\max(0, \sum_{k=1}^{n} Z_{k,j} - V_j))^2$$

Inspired from [Subramanian2017]

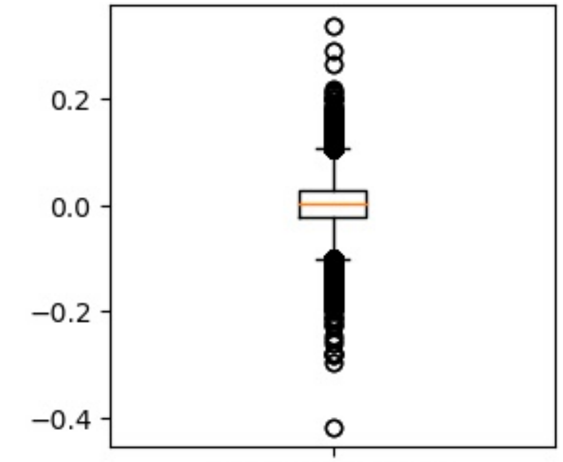
☞ ***Regulate the latent space during training pushing speaker profiles to respect a desired typicality of attributes.***

$$Loss = MSE + \lambda * L_S$$

Reminder: The presence of attribute in one utterance → Its presence in the profile.

ASpR performance

- Small correlation between BAE attributes.



Correlation values between BAE attributes

ASpR performance of BAE auto-encoder and the BA-extractor on VoxCeleb1

	BAE auto-encoder				Baseline
	Input	Latent space		Output	BA-extractor
Vector	Xvector	Z	BAE-vectors	$\widehat{Xvector}$	BA-vectors
#Dimensions	256	512	512	256	205
Evaluation	Cosine	Cosine	BA-LR	Cosine	BA-LR
EER	1.37%	2.22%	**2.46%**	1.8%	3.5%

☞ In terms of reconstruction, an increase of 0.43% in EER compared to the input.

☞ Compared to the x-vectors, an absolute increase of **only** ~1% in EER with BA-LR scoring.

☞ Compared to BA-vectors, a **relative reduction** of 30% of the EER with BA-LR scoring.

Key takeaways

- Address the limitations of the initially proposed BA-extractor.
- A binary auto-encoder, BAE, that introduces a loss to guide the binary vectors toward the desired behavior of attributes.

✓ BAE vectors present attribute-like behavior.

✓ BAE improves significantly the ASpR performance using BA-LR scoring.

✓ The results are promising and highlight the high potential of BA-LR approach.

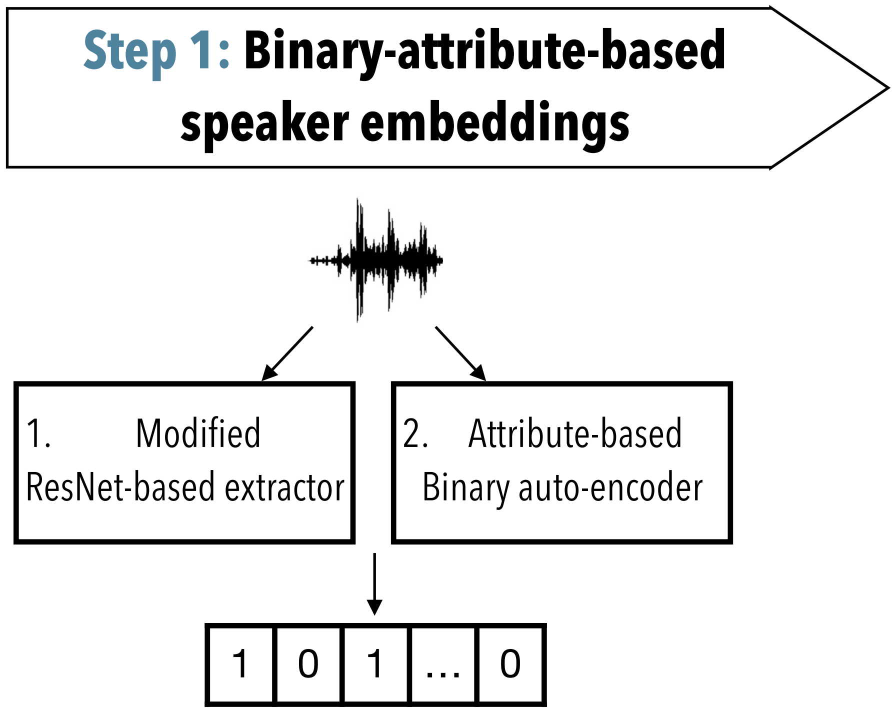
✗ The input x-vectors are not the best.

✗ The BAE model needs to be improved.

Conclusion & Perspectives

Conclusion

RQ1: Can we make the embedding space interpretable?

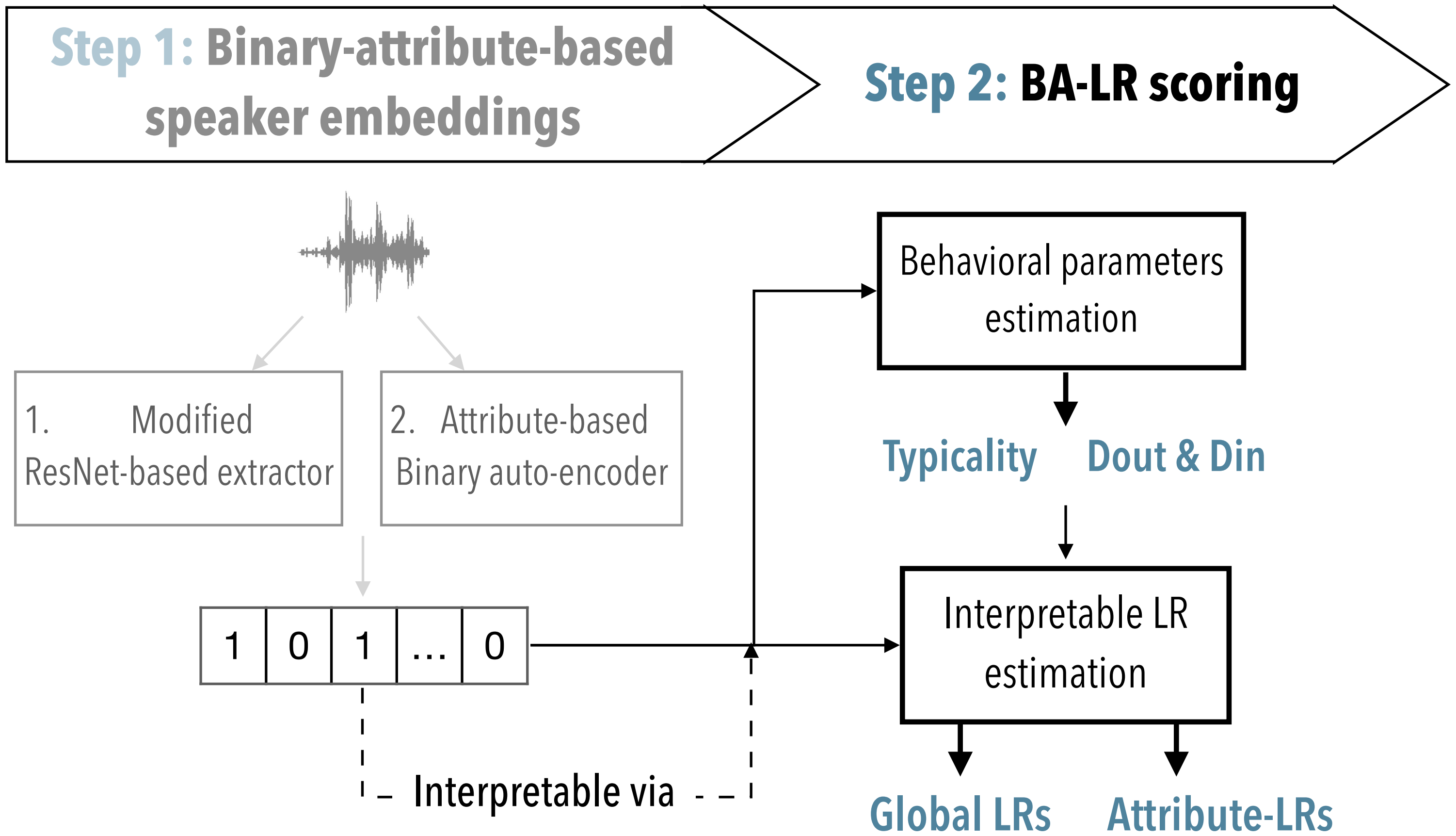


☞ Slight loss in performance compared to SOTA ASpR system.

☞ Easy to understand, simple restructuring of speaker information.

Conclusion

RQ2: Which voice information influences the final score in ASpR task? what is its contribution? Is it reliable?



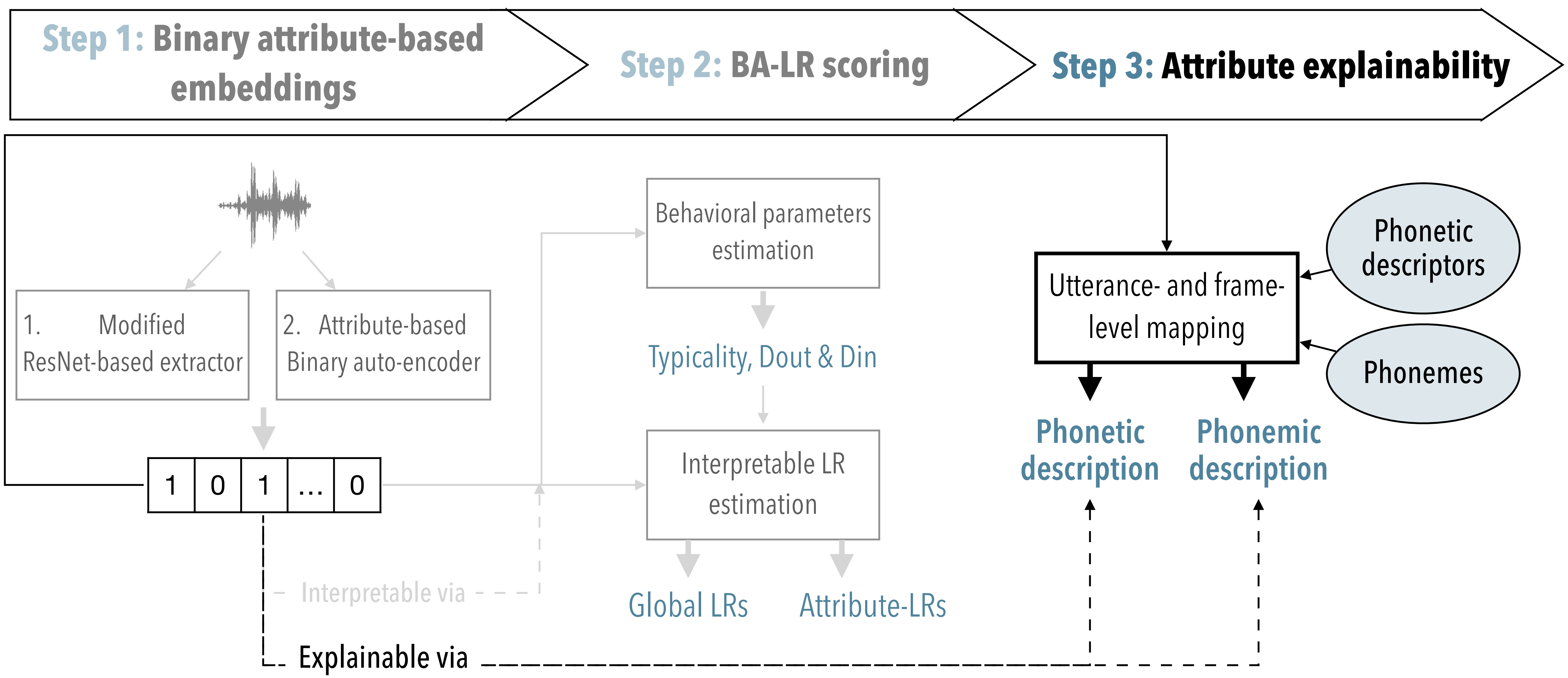
☞ Attributes are interpretable by their behavior and contribution.

☞ Present a transparent LR computation driven by the contribution of discriminant attributes.

☞ Good ASpR performance and generalisation abilities with BA-LR scoring.

Conclusion

RQ3: What is the nature of this encoded information?



☞ Offers insights about voice information encoded and involved into the ASpR scoring.

☞ Discovers phonetic combinations that encode high level features.

Conclusion

An **application** of BA-LR scoring on forensically realistic data is performed for validation.

☞ Generalisation ability of BA-LR on Dutch dataset.

☞ The weighted fusion of attribute-LLRs improved BA-LR scoring.

☞ This thesis opens a new perspective on explainable and interpretable ASpR systems.

☞ A helpful tool to understand information encoded by DNN models and aid for the court in making informed decisions.

☞ Its applicability extends far beyond forensic scenarios.

Perspectives

- Fine-tuning the BA-extractor with the attribute-based loss and STE technique to directly obtain binary speaker embeddings.
- The independence assumption between attributes might be involved as a constraint during training.
- Application of BA-LR approach on language or emotion identification.
- Beneficial to hide and better handle particular voice attributes for a privacy-related task.
- A suggestion of applying BA-LR on other types of data like forensic text comparison [Ishihara2020].

Related personal publications

- **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" Submitted in **Interspeech 2024.**
- **Imen Ben-Amor**, Jean-François Bonastre, Benjamin'O Brien, Pierre-michel Bousquet, "Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition",In: **Interspeech 2023**
- **Best Paper Award: Imen Ben Amor** and Jean-François Bonastre, "BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison," In: **IWBF2022.**
- **Imen Ben Amor** and Jean-François Bonastre, Abstract submission in **EAFS 2022** abstract book p 229.
- **Imen Ben-Amor** and Jean-François Bonastre. " BA-LR : une approche transparente de comparaison de voix en criminalistique". In: **JEP 2022**.

Best Paper Award in IWBF
Salzburg, Austria 2022

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**.
- Marie Tahon, Imen Ben-Amor et.al. "Interpretabilité pour l'identification de locuteurs. Retour sur le projet JSALT 2023", Journée commune **AFIA-TLH / AFCP 2023**.

Participation in international JSALT workshop 2023, LeMans

Thank you

Imen Ben-Amor

Imen.ben-amor@univ-avignon.fr

BA-LR Github code

References

- Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. “A time delay neural network architecture for efficient modeling of long temporal contexts”. In: Interspeech 2015.
- Yang Zhang et al. MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification. 2022. arXiv.
- 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.
- Sergey Novoselov et al. “Robust Speaker Recognition with Transformers Using wav2vec2.0”. In: ArXiv 2022
- Th. Kirat et.al “Fairness and explainability in automatic decision-making systems. A challenge for computer science and law” EURO journal on decision processes 2023.
- Abiodun A.Solanke “Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models” Forensic science international 2022.
- Ashley S.Deeks “The judicial demand for explainable artificial intelligence” Law & Society 2019
- Hossein Zeinali et al. “BUT System Description to VoxCeleb Speaker Recognition Challenge 2019”. In: arXiv:1910.12592. 2019.
- 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: SLT2018
- Margit Antal and Gavril Toderean. “Speaker Recognition and Broad Phonetic Groups”.in: Signal Processing, Pattern Recognition, and Applications. 2006
- Elie Khoury et.al “The 2013 Speaker Recognition Evaluation in Mobile Environment” ICB-2013
- Wiebke Toussaint Hutiri, Aaron Yi Ding“Bias in Automated Speaker Recognition”
- W. Hutiri, L. Gorce, and A. Y. Ding. “Design Guidelines for Inclusive Speaker Verification Evaluation Datasets”, Interspeech 2022
- Petros Boufounos and Shantanu Rane. “Secure binary embeddings for privacy preserving nearest neighbors”.IEEE International Workshop on Information Forensics and Security. 2011
- Lantian Li et al. “Binary speaker embedding”. ISCSLP2016
- Jean-Francois Bonastre et al. “Speaker modeling using local binary decisions”. In:Proc. Interspeech2011.