Interprétabilité pour l'identification de locuteurs

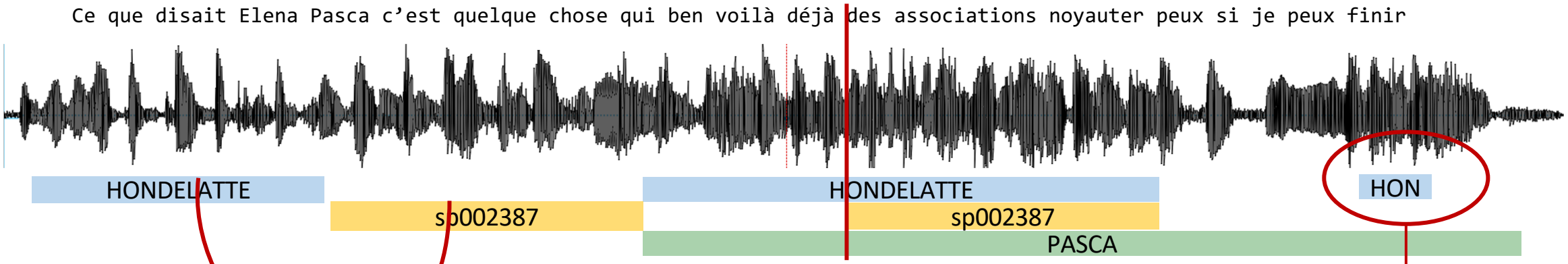
M. Tahon, I. Ben Amor, N. Dugué, J.F. Bonastre

Retour sur JSALT 2023

LIUM
Laboratoire d'Informa
Le Mans Université

Le Mans Université

Contexte : projet XDiar



Two different speakers

Q why are they different?

⇒ left: male, right: female

⇒ F0 explains 70% of the decision

The boundary is in the middle of a word

Q why the model found a boundary here?

⇒ left: no music, right: music

⇒ presence of break in linguistic cohesion

Low confidence

Q why this segment has been clustered with the other blues ?

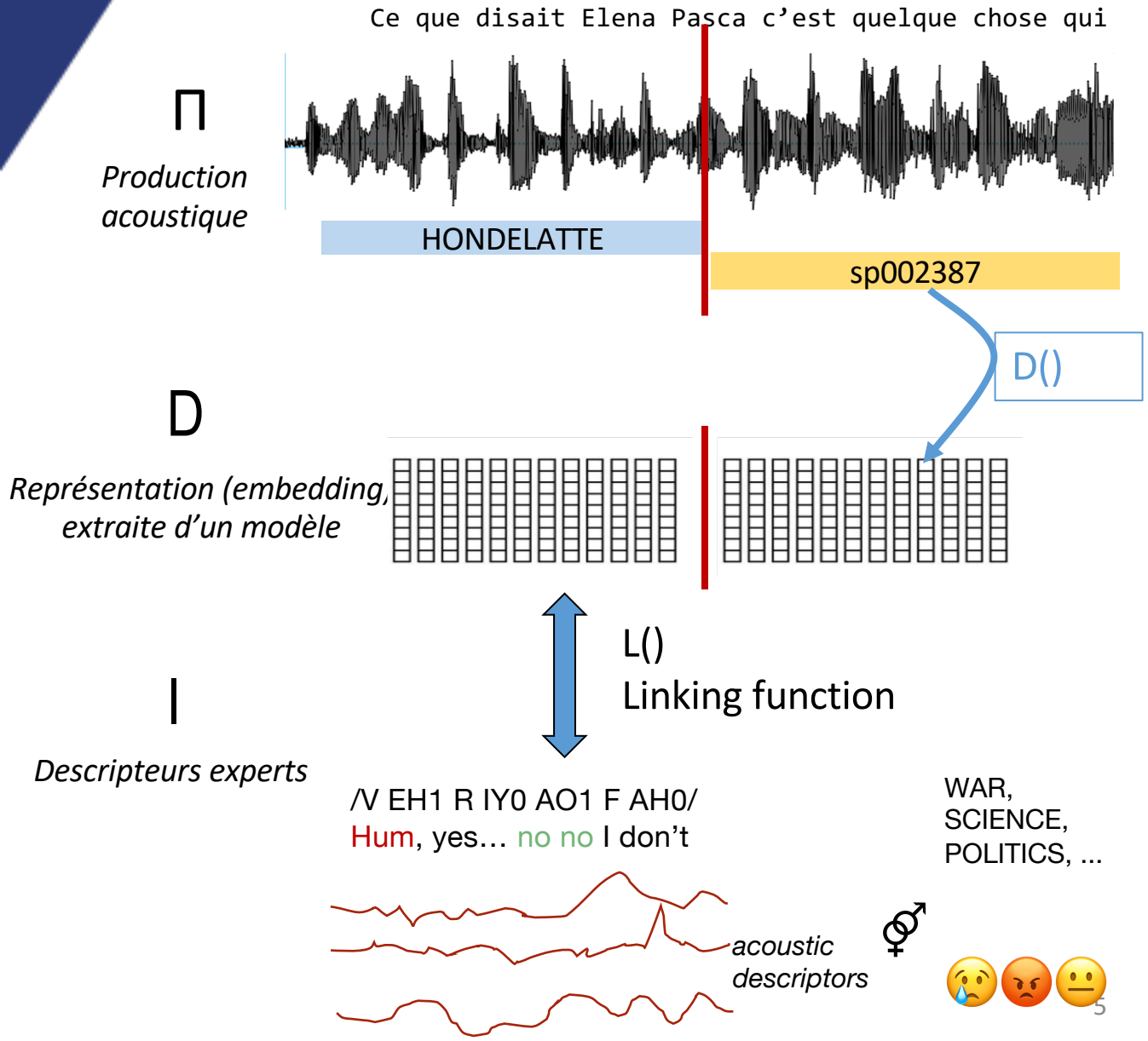
⇒ emotion: high activation

⇒ loudness explains 64% of the decision

Traitement automatique de signaux multi-locuteurs (partie I)

Expliquer les décisions (partie II)

Monde réel (Π) et monde des représentations (D)



Discrimination humaine

- le timbre,
- la prononciation,
- l'environnement acoustique

Expliquer les modèles

- Faire du probing sur le genre, l'émotion, les phonèmes [Ma 2021]
- Causalité (mécanismes de perturbation) [Lenglet 2022]

Interpréter les dimensions

- Multi-modalité
- Apport de connaissances expertes

Etat de l’art – aspects cognitifs

Des dimensions interprétables ?

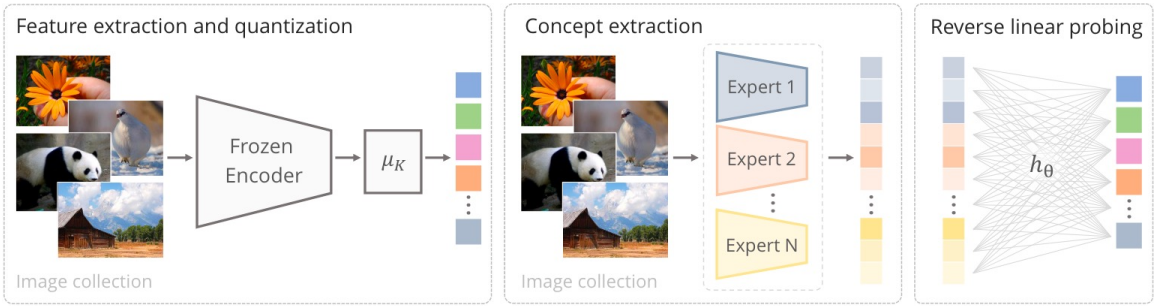
- Importance de la parcimonie
  - Caractéristiques typiques (circuit 24h)
- Positivité
  - Présence d’une caractéristique (circuit 24h/accent du sud)
  - Mieux que absence (Tour Eiffel/pas d’accent régional)
- Binaire
  - Pareil/différent ⇒ facile à interpréter

théorie de la Gestalt où la forme prime sur les parties

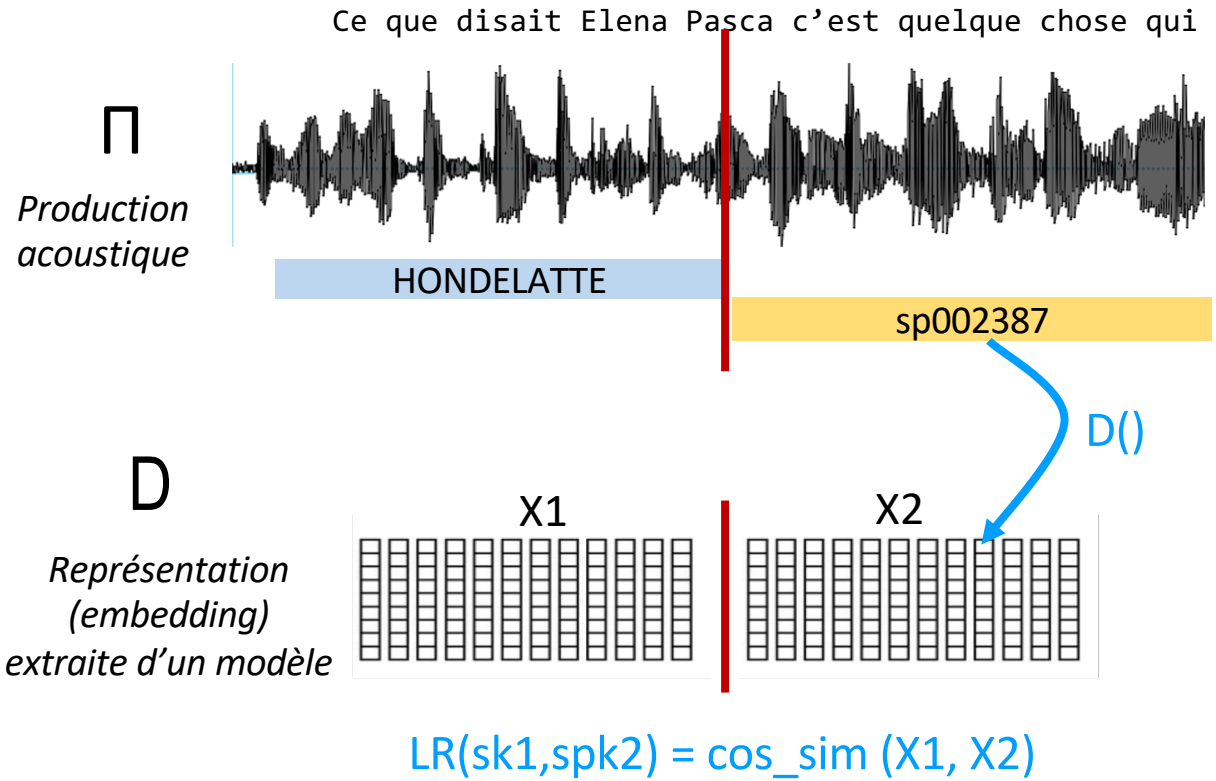
  - Pour certaines caractéristiques (bleu/genre) un continuum est préférable

Etat de l'art – interpréter les espaces de représentation

- Importance de parcimonie et la positivité
  - *SPINE: SParse Interpretable Neural Embeddings [Subramanian, AAAI 2018]*
- Importance de la binarisation pour l'interprétabilité
  - *BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison* [Ben Amor, IWBF 2022]
- Comment trouver le lien entre cet espace binaire et les descripteurs experts ?
  - *Measuring the interpretability of unsupervised representations via quantize reverse probing* [Laina, ICLR 2022]



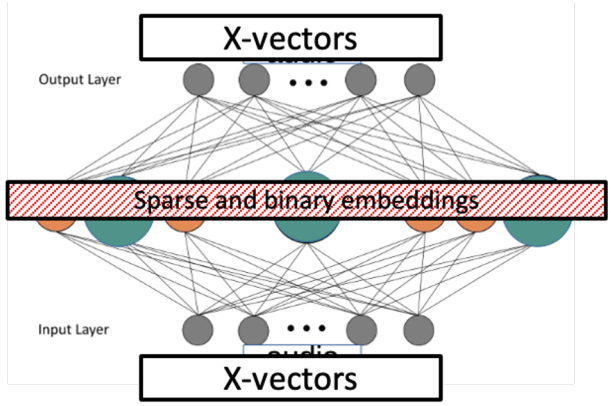
La vérification du locuteur



- Modèle identification du locuteur:
  - ResNet64
  - Appris sur VoxCeleb2
  - EER(vox1) = 1.37
- D() = x-vector
  - 256 dimensions
- Similarité entre x-vectors
  - Rejet/acceptation

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Construction d'un espace parcimonieux et presque binaire



256

500

256

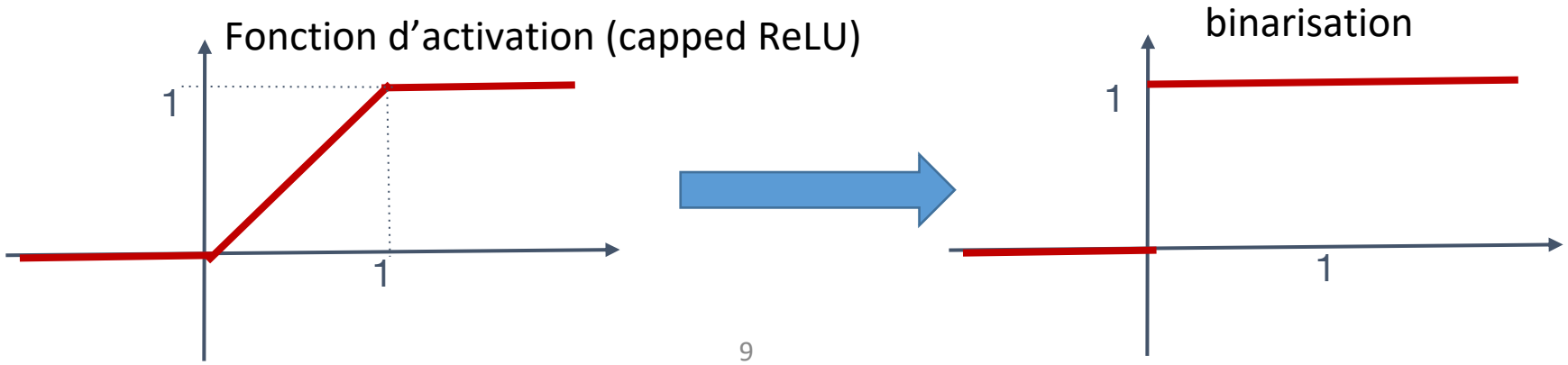
SPINE : auto-encoder à 1 couche linéaire

Reconstruction loss : $RL(D)=\frac{1}{|D|}\sum_{X\in D}\|X-\tilde{X}\|_2^2$

Partial Sparsity loss : $PSL(D)=\frac{1}{|D|}\sum_{X\in D}\sum_{h\in H}(Z_h^{(X)}*(1-Z_h^{(X)}))$

Average Sparsity loss : $ASL(D)=\sum_{h\in H}\max(0,p_{h,D}-p_{h,D}^d)^2$

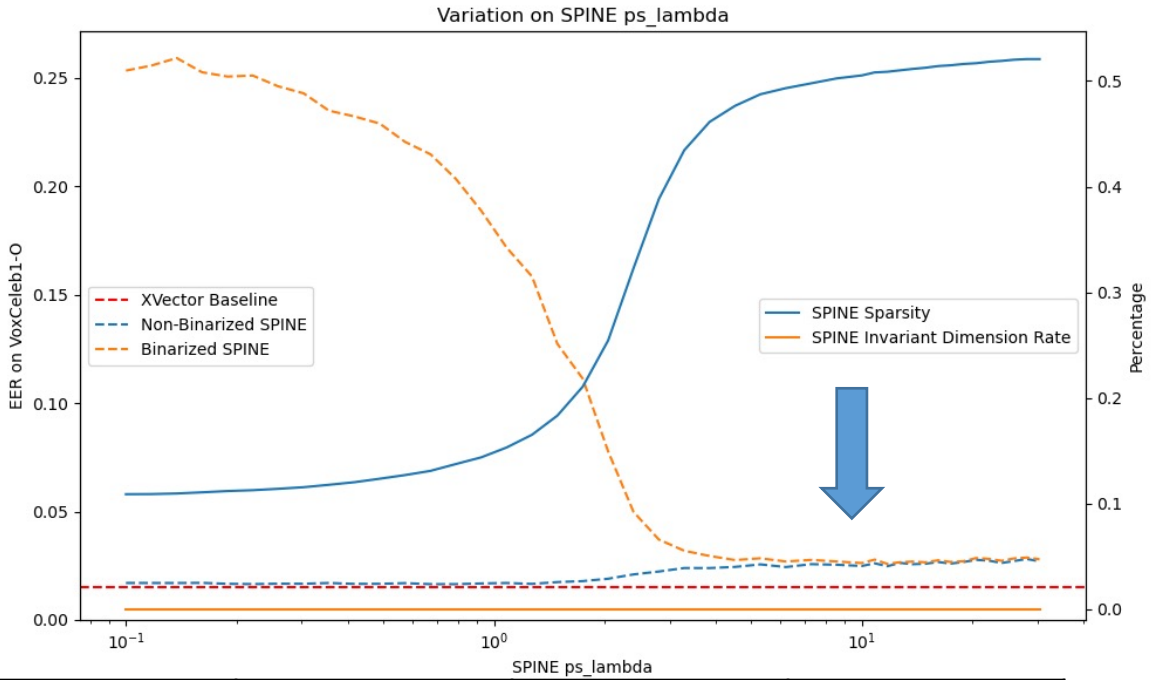
Final loss : λ_1RL(D) + λ_2PSL(D) + λ_3ASL(D)



Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., & Hovy, E. (2018). SPINE : SParse Interpretable Neural Embeddings. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1)

Représentations SPINE

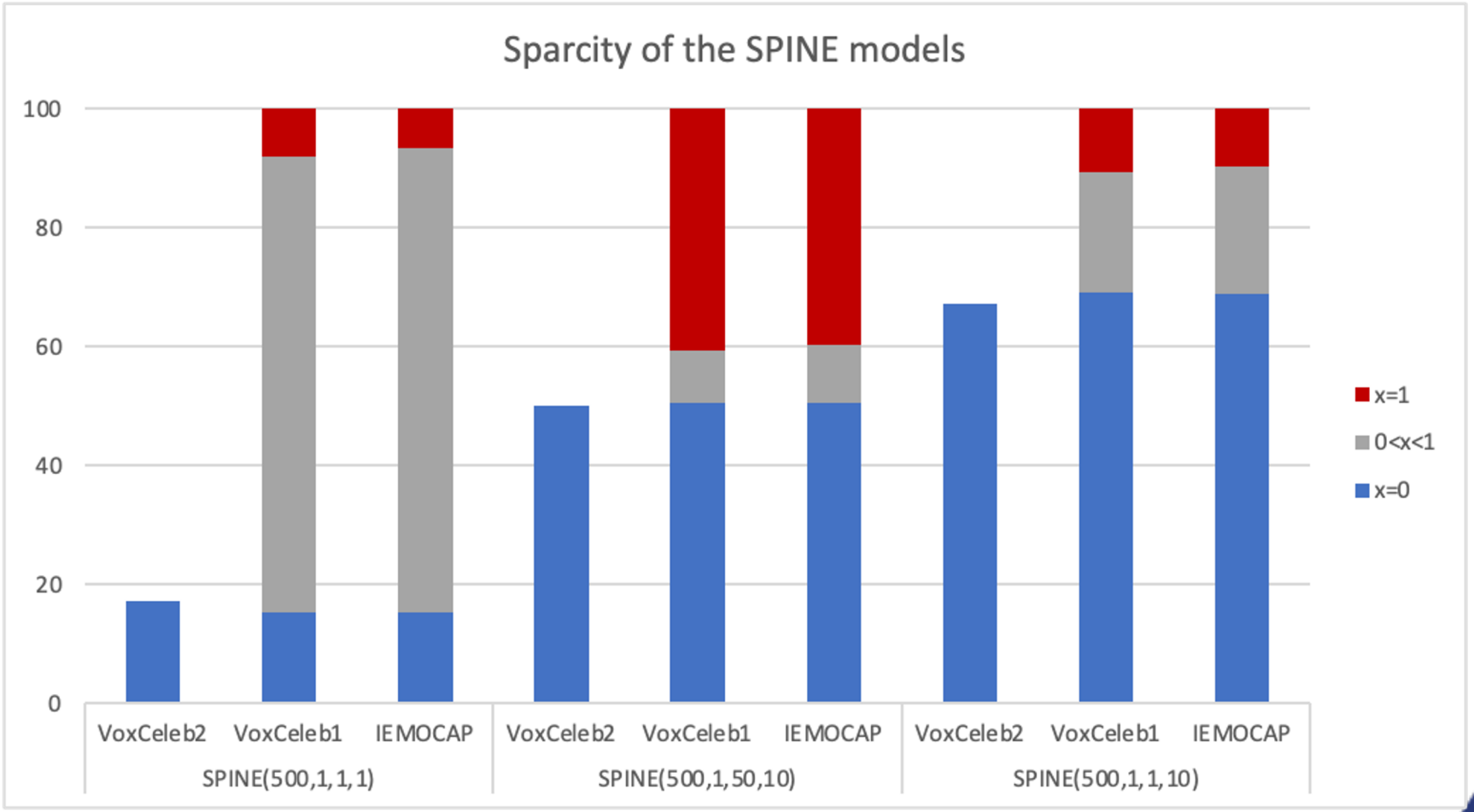
- Optimisation de 4 paramètres
  - Final loss : λ_1RL(D) + λ_2PSL(D) + λ_3ASL(D)
  - Dimension
- Choix d'un *bon* modèle
  - Parcimonie
  - Détection du locuteur EER(vox1)



R λ_1	ASL λ_3	PS λ_2	Input size	Embedding size	EER (Non Binarized)	EER (Binarized)	Sparsity
1	1	1	256	500	1.66%	1.76%	15%
1	1	10	256	500	2.41%	2.58%	50%
1	50	10	256	500	3.15%	3.25%	69%

Evaluation de la parcimonie

Reference:
x-vec
S = 0% (taux de valeurs nulles)
EER(vox1) = 1.37



S = 15%
EER(vox1) = 1.66

S = 50%
EER(vox1) = 2.41

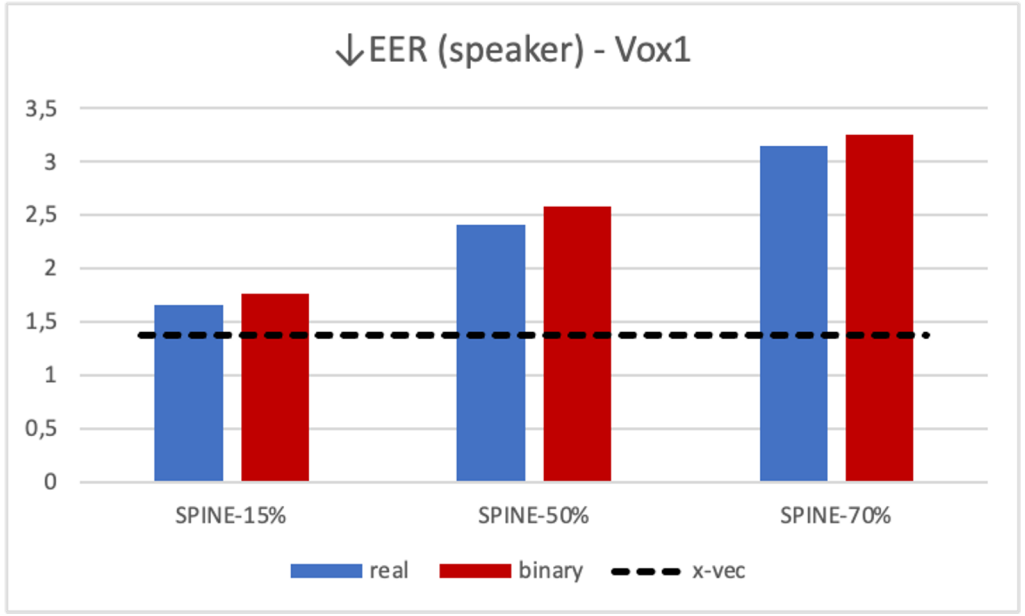
S = 70%
EER(vox1) = 3.15

Evaluation

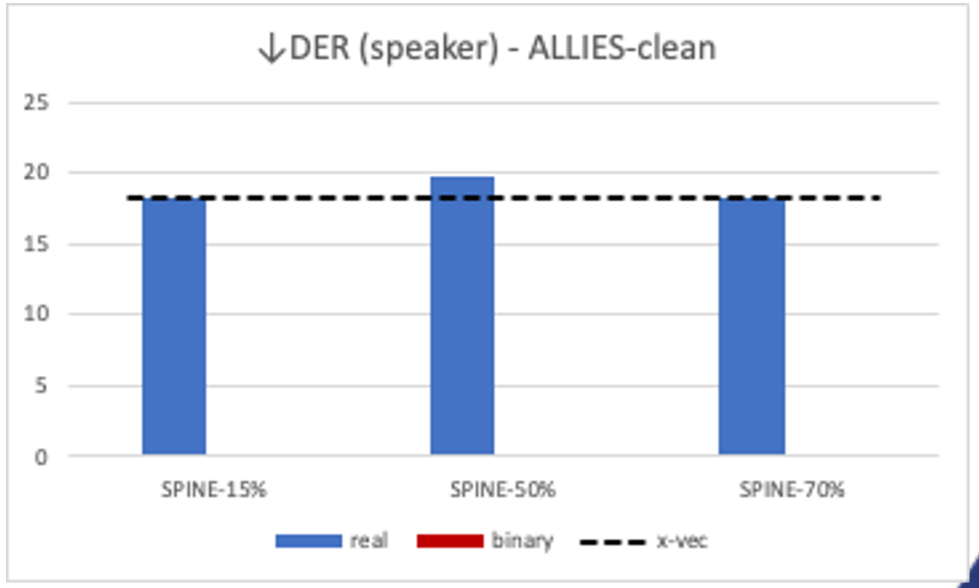
Identification du locuteur EER(vox1), diarization (DER)

- Parcimonie ↗ EER mais maintient DER
- Binarisation ↗ (un peu) EER

⇒ SPINE-70% contient presque que des 0, et perd seulement 1.9p



x-vector + oracle VAD + VBx

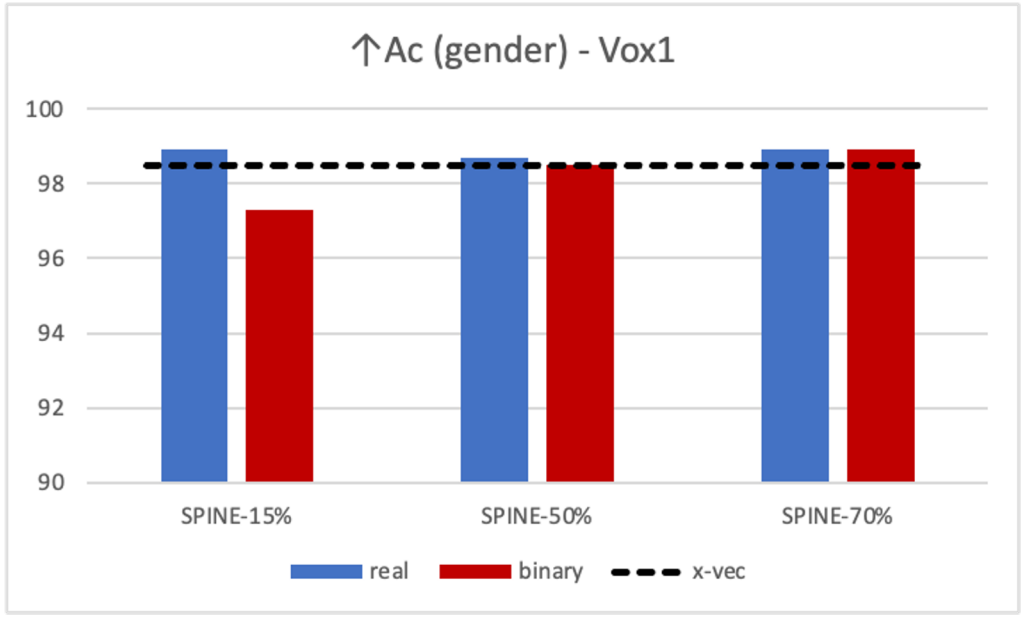


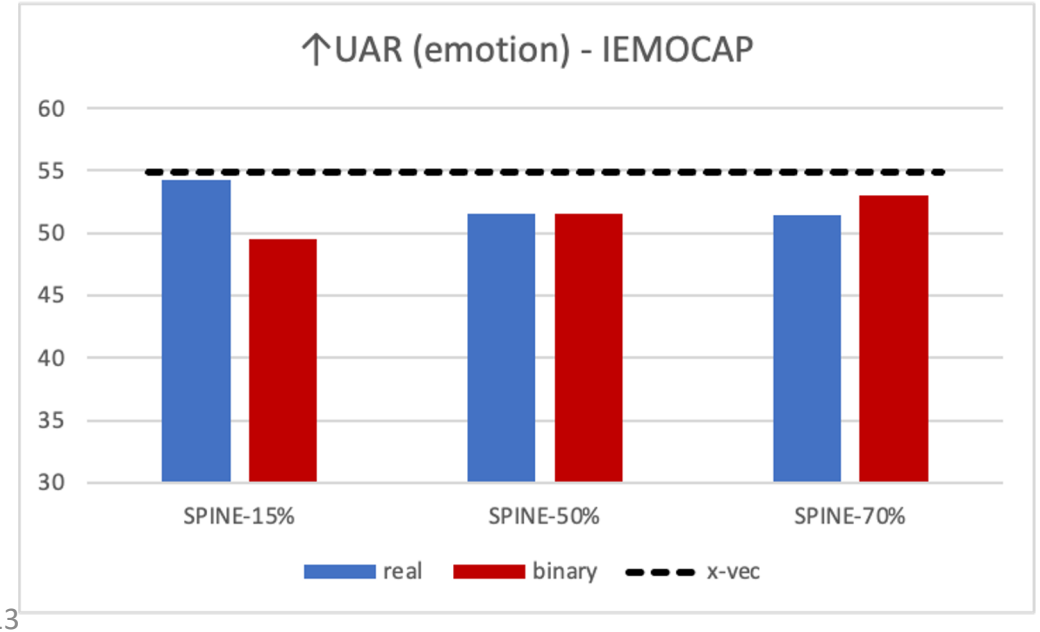
Evaluation

Classification du genre (Acc), des émotions – 4 classes (UAR):

- Parcimonie : ↗ Genre, ↘ Emotion
- Binarisation : ↘ avec SPINE-15% mais ↗ with SPINE-70%

⇒ Les SPINE-70% binaires contiennent presque toute l'information





Déterminer les dimensions *typiques* de D

Objectif : explorer l'espace de représentation D (vecteurs SPINE) et chercher des dimensions typiques.

2 classifieurs: genre et émotion

Sélection des dimensions les plus typiques:

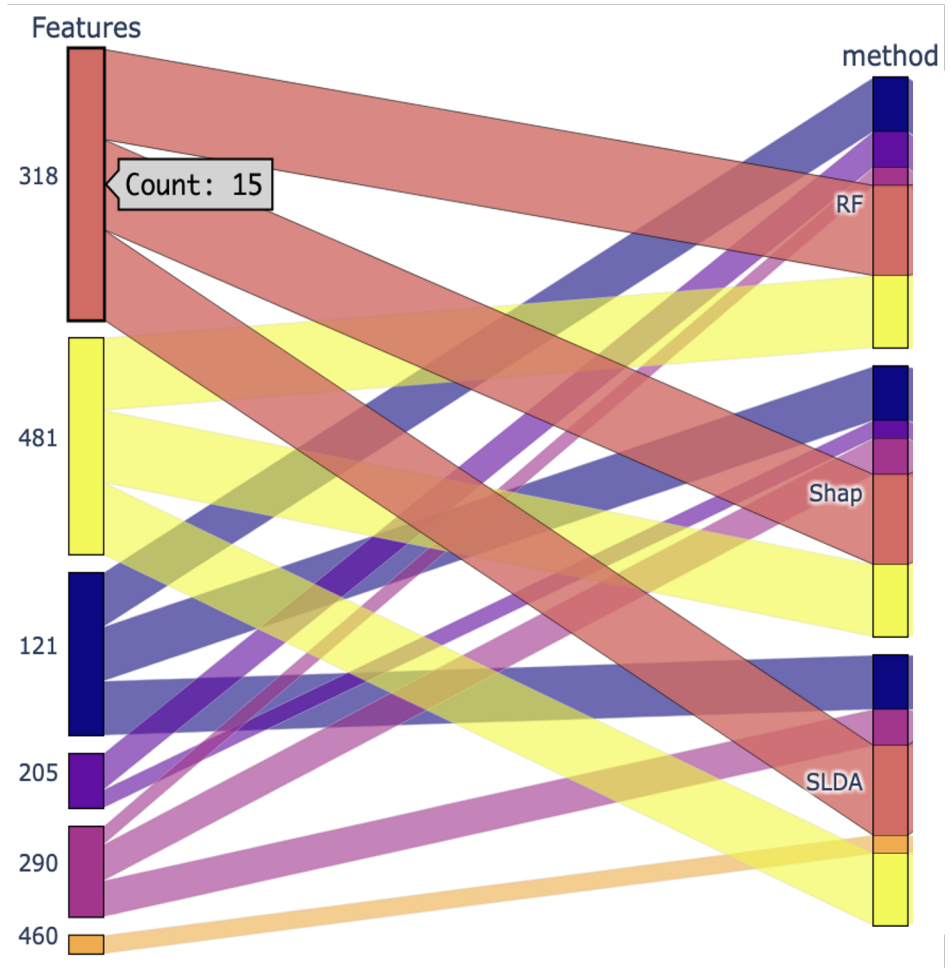
- Classifieur + ordre d' importance
- Classifieur + valeurs de Shapley
- LDA-based discriminant selection (SLDA)

Dimensions typiques SPINE-15% – genre

- Classifieur : RandomForest
- Test : Vox1
- Avec les 500 dimensions Ac=99%
- Tri des dimensions par ordre d'importance
  - Critère de Gini
- Avec les 4 dimensions les plus importantes ⇒ Ac = 75%
- Résultats similaires obtenus avec SLDA et valeurs de Shapley.

Les 6 dimensions les plus importantes

La taille de la ligne correspond à l'importance de la dimension

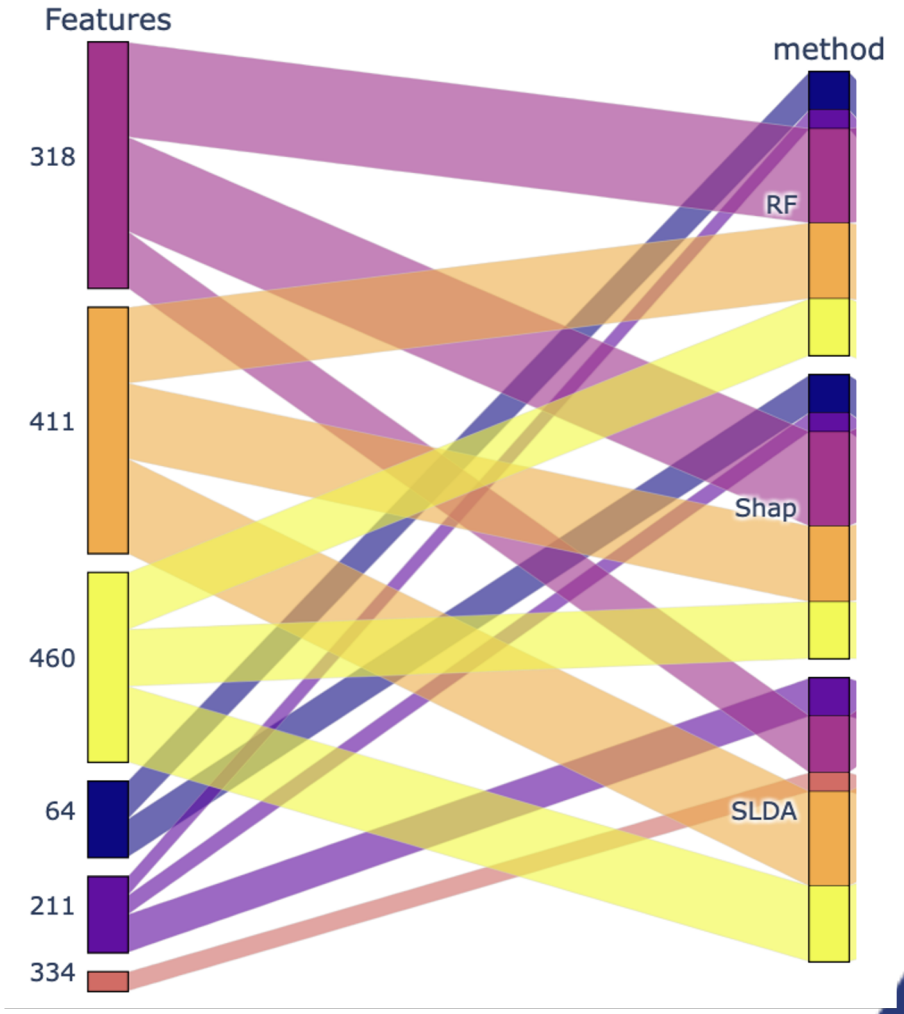


Dimensions typiques SPINE-70% – genre

- Classifieur : RandomForest
- Test : Vox1
- Avec les 500 dimensions Ac=98%
- Tri des dimensions par ordre d'importance
  - Critère de Gini
- Avec les 4 dimensions les plus importantes ⇒ Ac = 75%
- Résultats similaires obtenus avec SLDA et valeurs de Shapley.

Les 5 dimensions les plus importantes

La taille de la ligne correspond à l'importance de la dimension



Déterminer la relation entre D et I

Objectif: Lier l'espace de représentation (D) aux descripteurs experts (I)

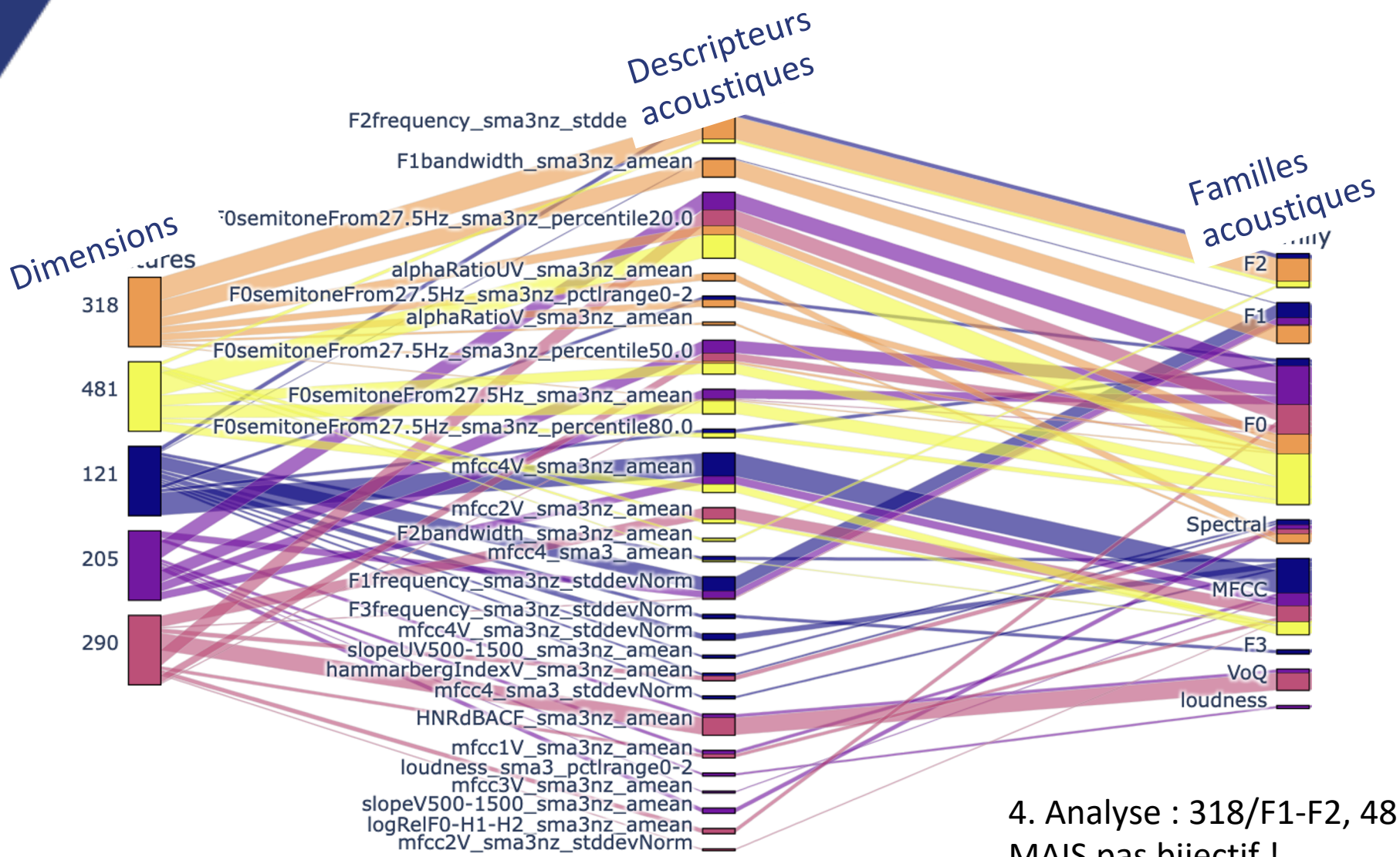
- D = $\{d_1, d_2, \ldots d_{500}\}$: vecteurs SPINE binaires
- I = $\{i_1, i_2, \ldots i_{88}\}$: descripteurs acoustiques (eGeMAPS)

- Analyses statistiques pour trier les descripteurs vs la dimension d_n
  - Correlation (pearson) entre dimensions binaires et descripteurs
  - Importance individuelle de chaque descripteur pour prédire chaque dimension (randomForest + Shapley/Gini)
  - LDA-based discriminant selection (SLDA)
  - Importance de paires de descripteurs basée sur l'information mutuelle – Double Input Symetrical Relevance (Disr)

$$J_{disr}(i_k) = \sum_j \frac{I(i_k, i_j | d_n)}{H(i_k, i_j, d_n)}$$

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

Relation entre D et I - genre



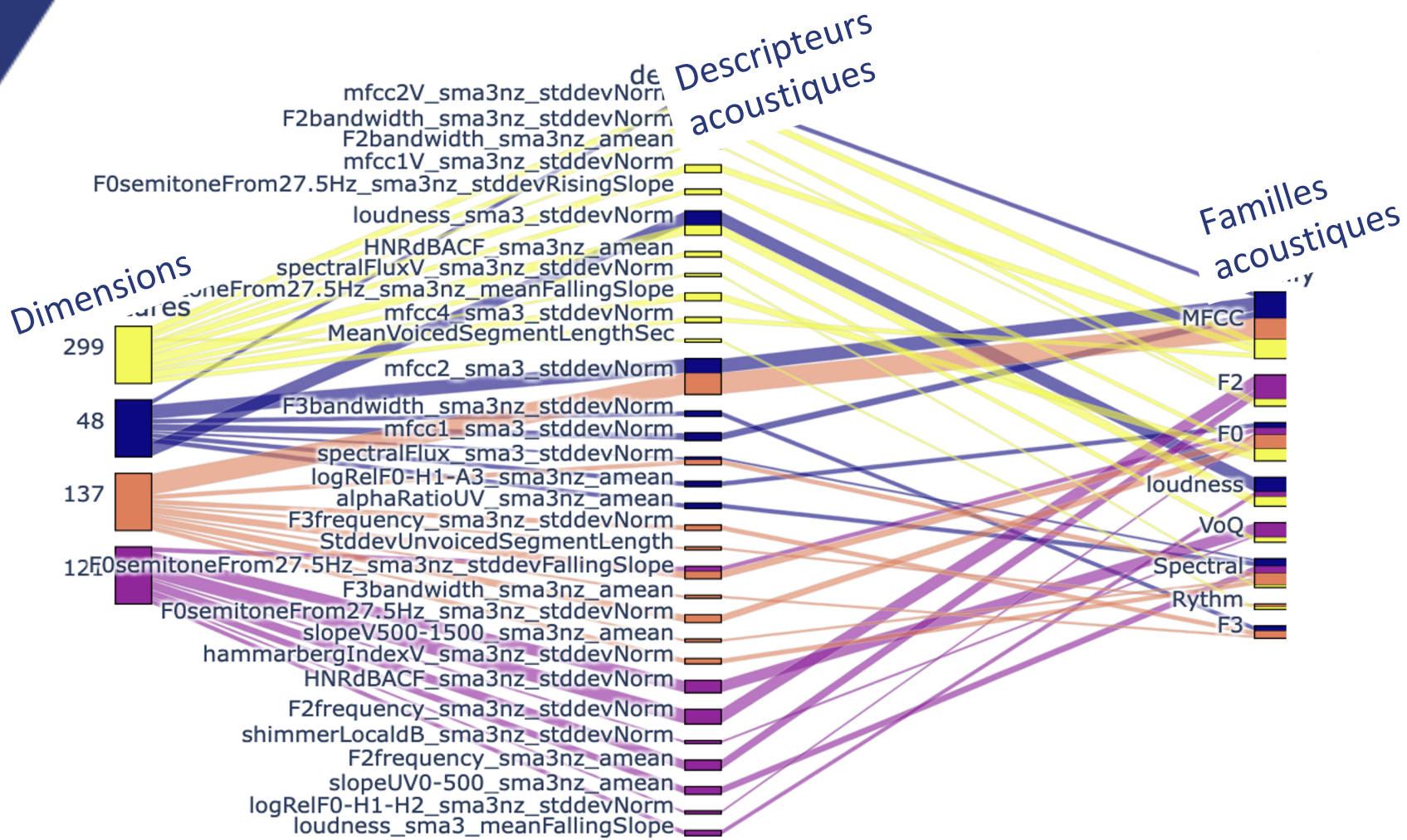
1. Prendre les 5 dimensions les plus typiques de SPINE – 15%

2. Trier les descripteurs acoustique par rapport à ces 5 dimensions prises individuellement (vox1) suivant les 4 méthodes

3. Calculer une valeur d'importance pour chaque descripteur en combinant les rangs obtenus par les 4 méthodes
⇒ épaisseur du trait

4. Analyse : 318/F1-F2, 481/F0, 121/MFCC, 205/F0
MAIS pas bijectif !

Relation entre D et I - émotion



1. Prendre les 5 dimensions les plus typiques de SPINE – 15%

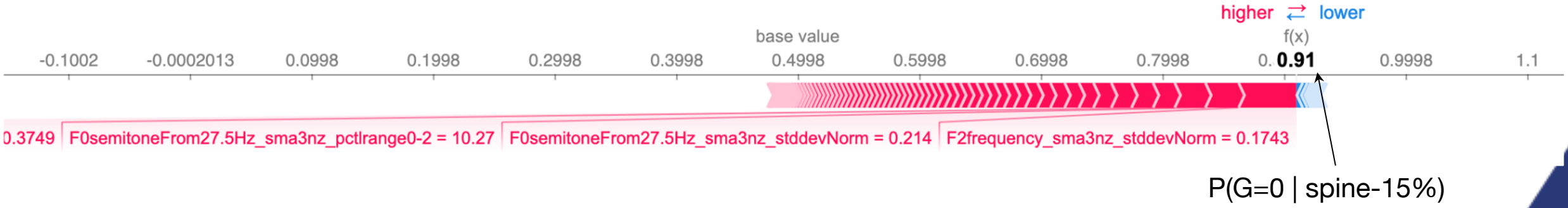
2. Trier les descripteurs acoustique par rapport à ces 5 dimensions prises individuellement (iemocap) suivant les 4 méthodes

3. Calculer une valeur d'importance pour chaque descripteur en combinant les rangs obtenus par les 4 méthodes
⇒ épaisseur du trait

4. Analyse : pas évident
Importance de la prosodie et du contenu spectral

Relation locale entre D et I avec Shap

- Expliquer localement quels descripteurs acoustiques (I) ont une influence sur la dimension 318 (D)
- 318 : dimension la plus importante pour prédire le genre
- Exemple réel de VoxCeleb1 *(id10481-r_2ZsMFY0fg-00003.wav)*
  - Prédiction pour la classe 0 (homme)
  - Descripteurs importants: variation de F2 sur le segment + F0



P(G=0 | spine-15%)

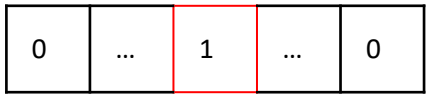
Application à la diarization

Π

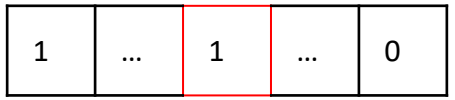
Production acoustique

D

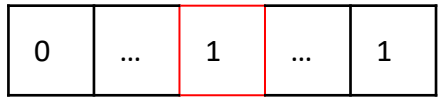
Représentation (embedding) extraite du modèle spine-15%



$d_{481}=1$



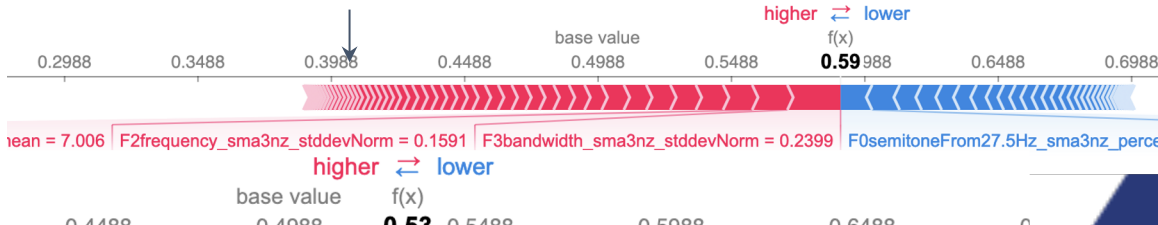
$d_{481}=1$

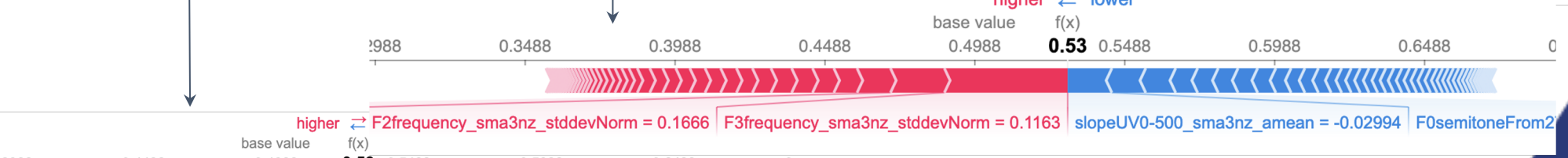


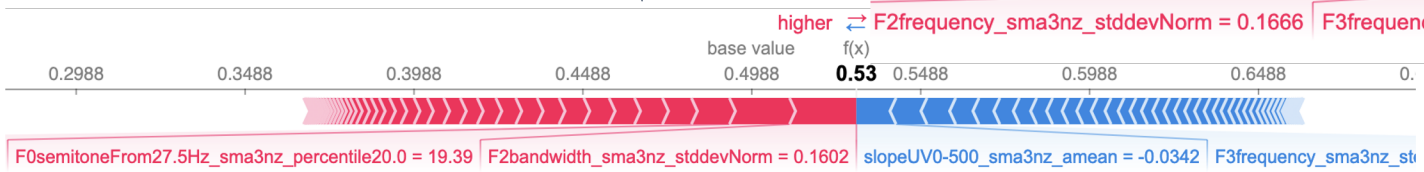
$d_{481}=1$

I

Descripteurs experts







Conclusions et perspectives

- Méthodologie pour l'interprétabilité en vérification du locuteur
  - Importance de la parcimonie
  - Importance de la positivité et de la binarisation
- Propositions pour lier l'espace de représentation aux descripteurs
  - Prendre en compte la dépendance entre les descripteurs ET entre les dimensions
  - Redéfinir l'ensemble de descripteurs (prosodie, phonétique, linguistique)
- Opération inverse :
  - modifier les descripteurs et voir l'effet sur la vérification