Segmentation audio explicable

Distillation de connaissances et factorisation matricielle non-négative

Théo Mariotte, Antonio Almudèvar, Marie Tahon, Alfonso Ortega

LIUM
Laboratoire d'Informatique
Le Mans Université

Universidad Zaragoza
1542

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666.

Contexte

Traitement automatique des archives audiovisuelles

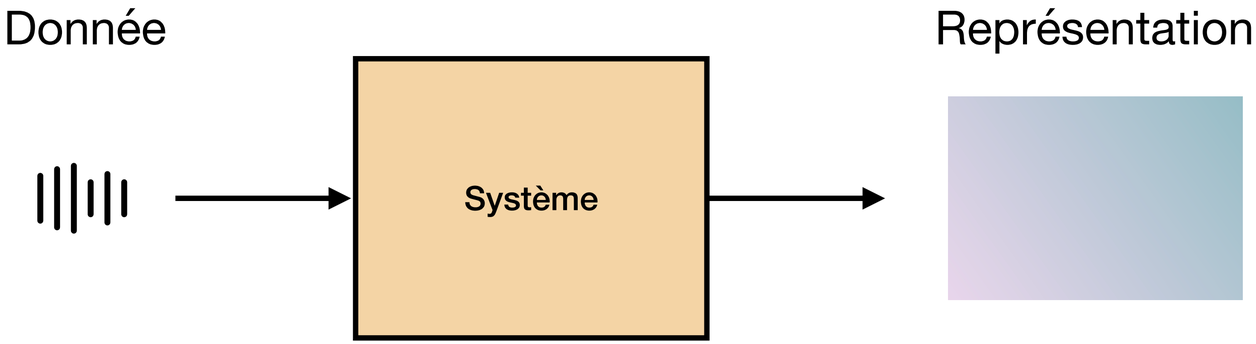
Source: https://arcmc.hypotheses.org/2460

- Grandes bases de données audiovisuelles
- Développement de méthodes de traitement automatique
- Parole : diarization en locuteurs, transcription automatique, reconnaissance du locuteur…
- État de l'art : réseaux de neurones artificiels
- **Problème** : difficile de comprendre l'information utilisée par ces systèmes

⇨ Besoin d'expliquer les décisions des systèmes !

Contexte

Expliquer oui… Mais comment ?

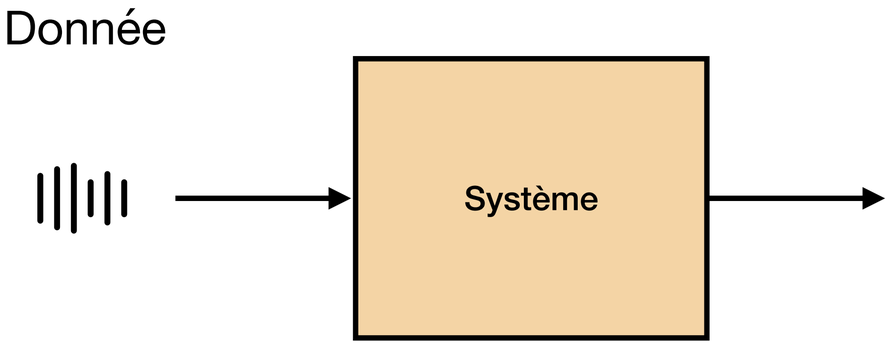


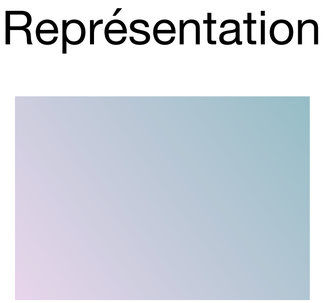
- Relation donnée/représentation abstraite
- Pas de sens « physique »

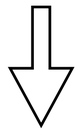
Subramanian, A. Et al. (2018, April). Spine: Sparse interpretable neural embeddings. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

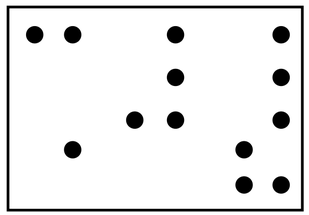
Contexte

Expliquer oui… Mais comment ?









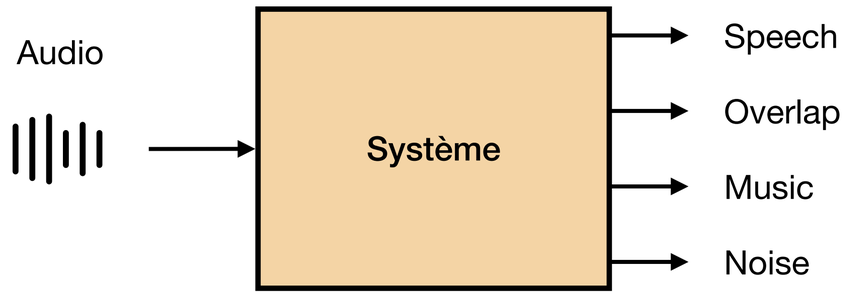
- Relation donnée/représentation abstraite
- Pas de sens « physique »

- Positivité
- Parcimonie
- Lien vers une représentation « signal »

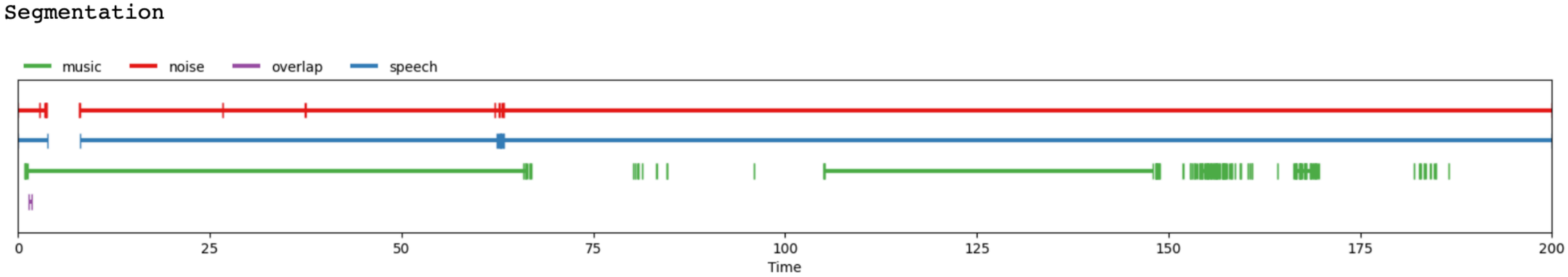
Subramanian, A. Et al. (2018, April). Spine: Sparse interpretable neural embeddings. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

Contexte

Segmentation automatique du signal audio



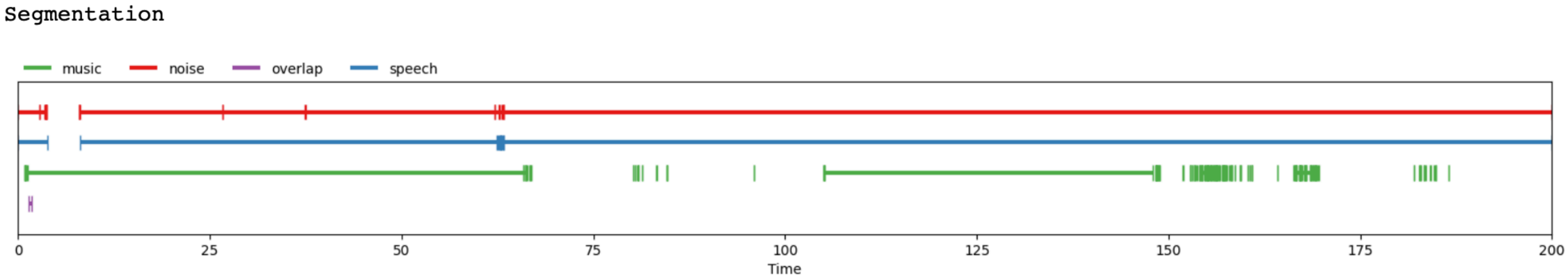
- Détection de la présence d'évènements en **fonction du temps**
- Utile à de nombreuses tâches de traitement automatique audio
- Évaluation à l'aide du F1-score



Contexte

Pourquoi segmenter ?

⇨ Utile à de nombreuses tâches de traitement audio



⇩ Speaker diarization

⇩ Music information retrieval

⇩ ASR

⇩ Analyse des erreurs

Contexte

Pourquoi expliquer la segmentation ?

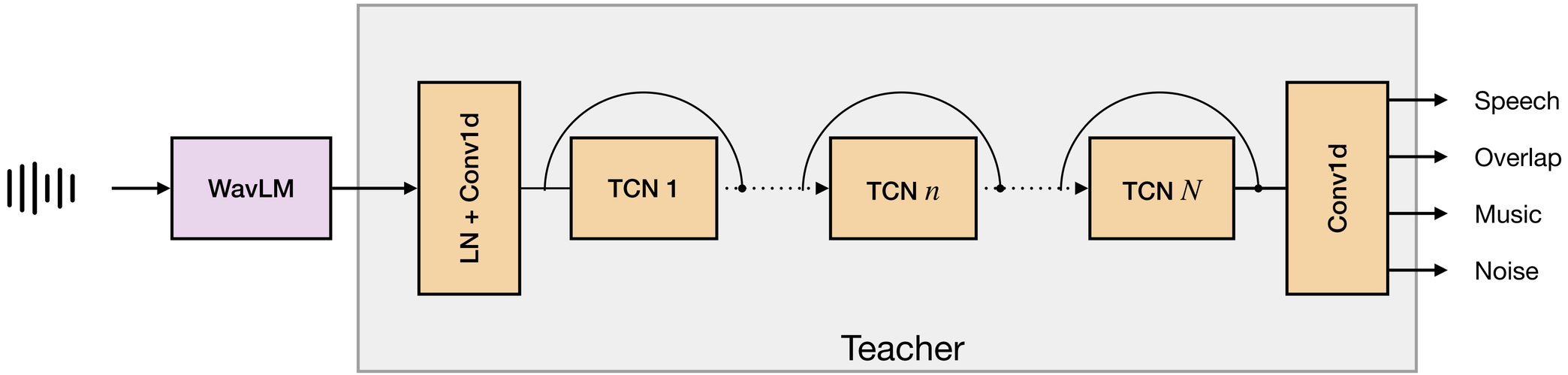
- Quelle information est utilisée dans le signal pour segmenter ?
- Comment sont discriminées des classes proches (parole/parole superposée) ?
- Peut on obtenir un prototype de classe ?

Sommaire

Sommaire

Segmentation audio explicable

Système *teacher*



- Représentation du signal audio avec WavLM (modèle auto-supervisé)
- Modélisation de séquence à l'aide d'un système TCN
- Prédiction des 4 classes à la trame : une prédiction à chaque instant

Chen, S. Et al. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505-1518.

Segmentation audio explicable

Factorisation matricielle non-négative

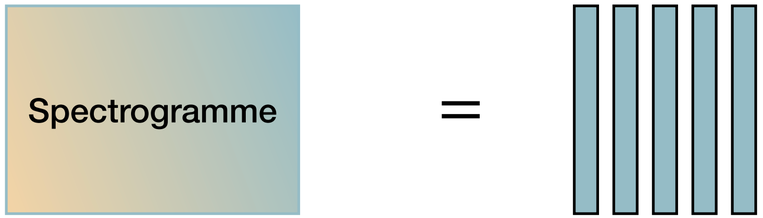
$$\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$$

Le Roux, J. Et al. (2015). Sparse NMF–half-baked or well done?. Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023, 11, 13-15.

Segmentation audio explicable

Factorisation matricielle non-négative

Dictionnaire



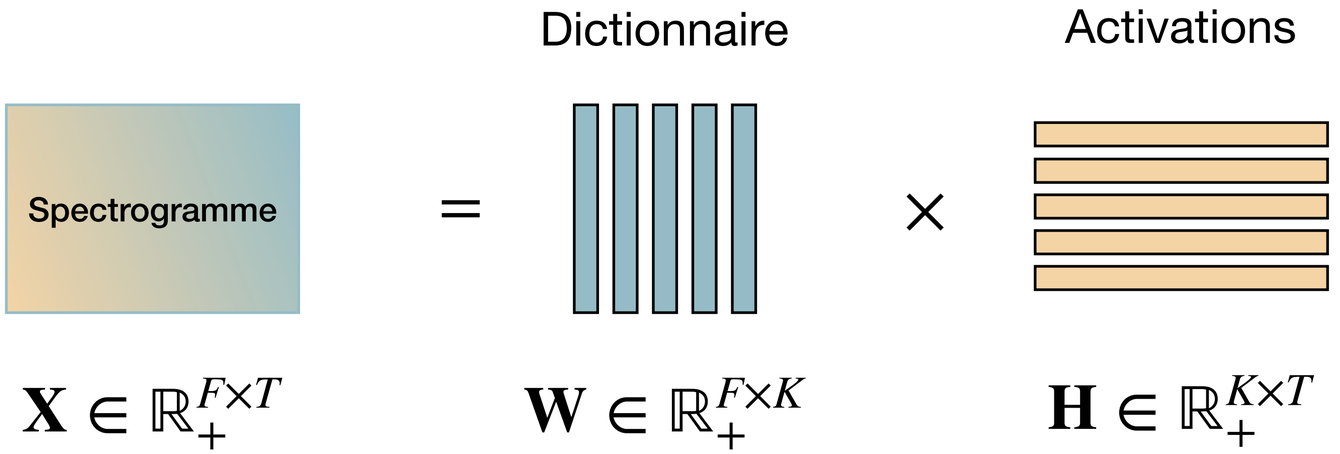
$\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$ $\mathbf{W} \in \mathbb{R}_{+}^{F \times K}$

« Fréquences »

Le Roux, J. Et al. (2015). Sparse NMF–half-baked or well done?. Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023, 11, 13-15.

Segmentation audio explicable

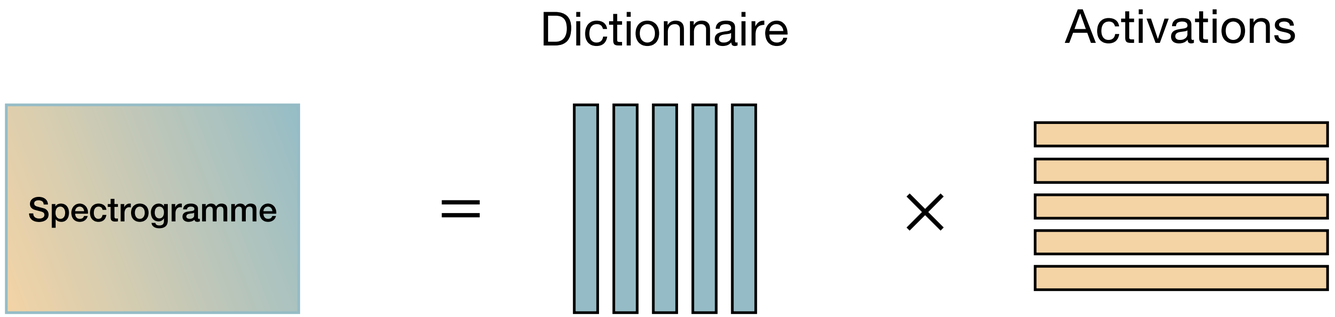
Factorisation matricielle non-négative



Le Roux, J. Et al. (2015). Sparse NMF–half-baked or well done?. Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023, 11, 13-15.

Segmentation audio explicable

Factorisation matricielle non-négative



$$\tilde{\mathbf{X}} = \mathbf{WH}$$

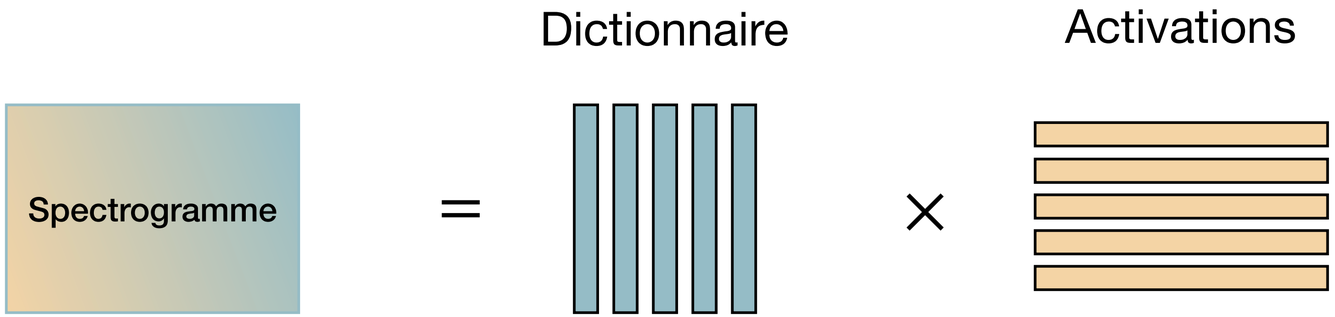
$\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$ $\mathbf{W} \in \mathbb{R}_{+}^{F \times K}$ $\mathbf{H} \in \mathbb{R}_{+}^{K \times T}$

« Fréquences » « Temps »

Le Roux, J. Et al. (2015). Sparse NMF–half-baked or well done?. Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023, 11, 13-15.

Segmentation audio explicable

Factorisation matricielle non-négative



$$\tilde{\mathbf{X}} = \mathbf{WH}$$

$$\mathbf{X} \in \mathbb{R}_{+}^{F \times T} \qquad \mathbf{W} \in \mathbb{R}_{+}^{F \times K} \qquad \mathbf{H} \in \mathbb{R}_{+}^{K \times T}$$

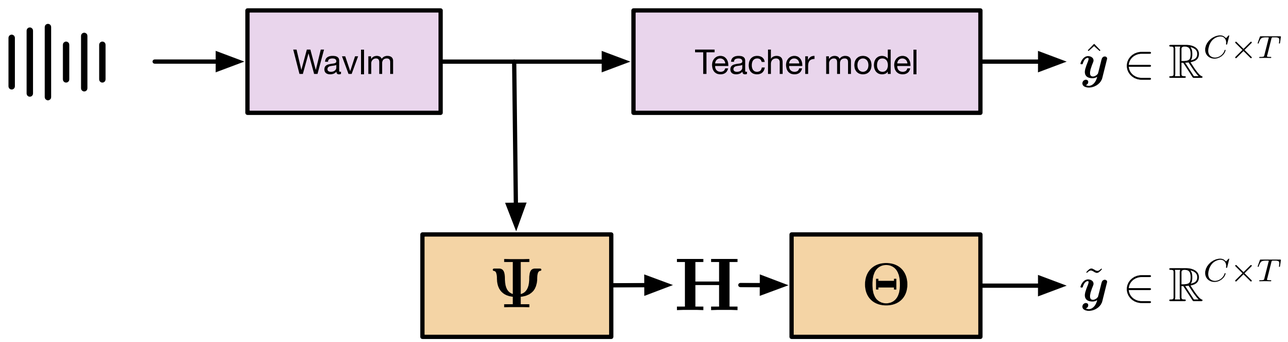
« Fréquences » « Temps »

NMF parcimonieuse $\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_2 + \mu\|\mathbf{H}\|_1$

Le Roux, J. Et al. (2015). Sparse NMF–half-baked or well done?. Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023, 11, 13-15.

Segmentation audio explicable

Architecture du *student*

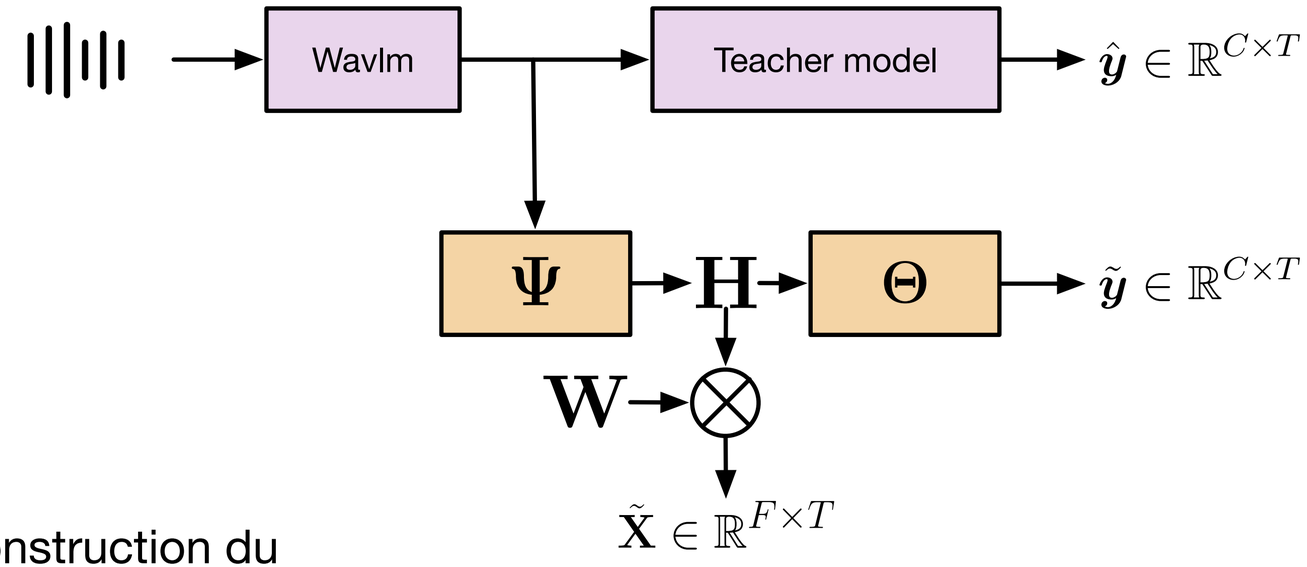


Objectif : contraindre l'espace de représentation $\mathbf{H}$ pour le rendre explicable

J. Parekh, et al. (2023) *Tackling Interpretability in Audio Classification Networks with Non-negative Matrix Factorization*. http://arxiv.org/abs/2305.07132

Segmentation audio explicable

Architecture du *student*



Contrainte 1: reconstruction du spectrogramme

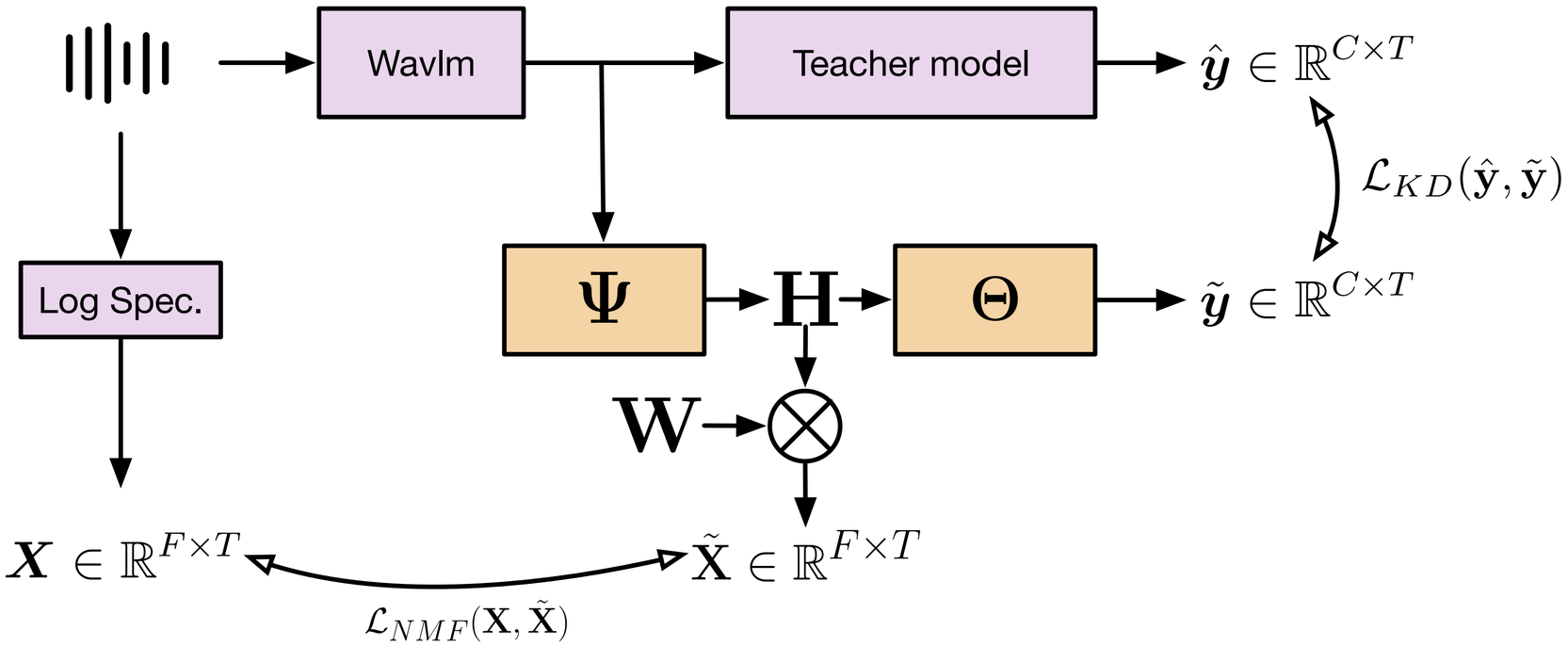
Contrainte 2: non-négativité

Contrainte 3: parcimonie

J. Parekh, et al. (2023) *Tackling Interpretability in Audio Classification Networks with Non-negative Matrix Factorization*. http://arxiv.org/abs/2305.07132

Segmentation audio explicable

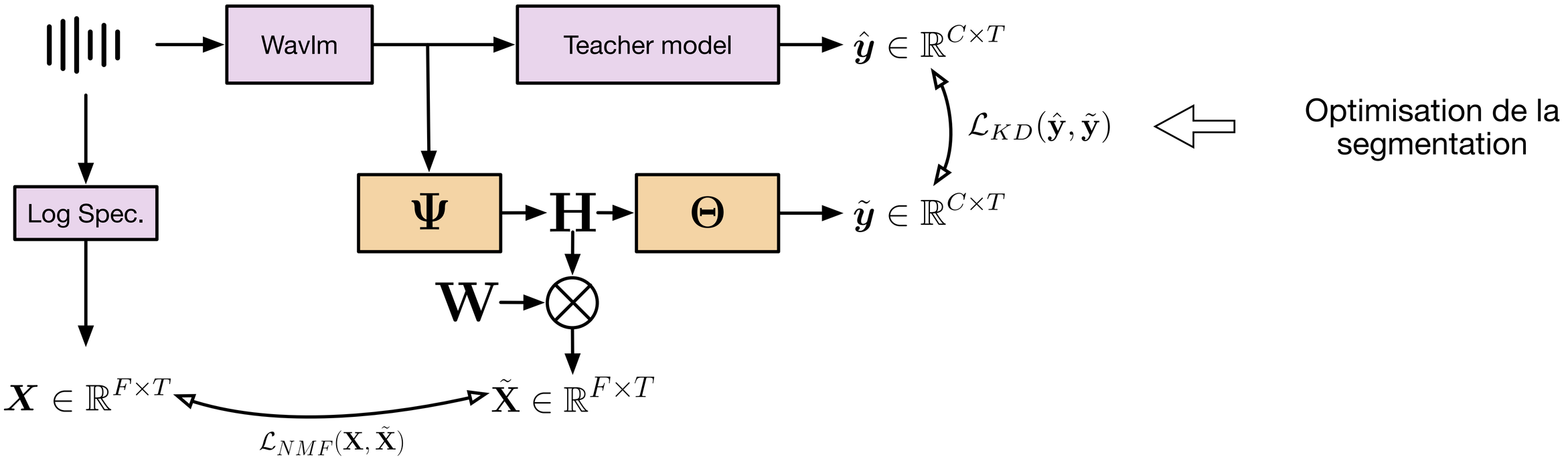
Distillation de connaissances



J. Parekh, et al. (2023) *Tackling Interpretability in Audio Classification Networks with Non-negative Matrix Factorization*. http://arxiv.org/abs/2305.07132

Segmentation audio explicable

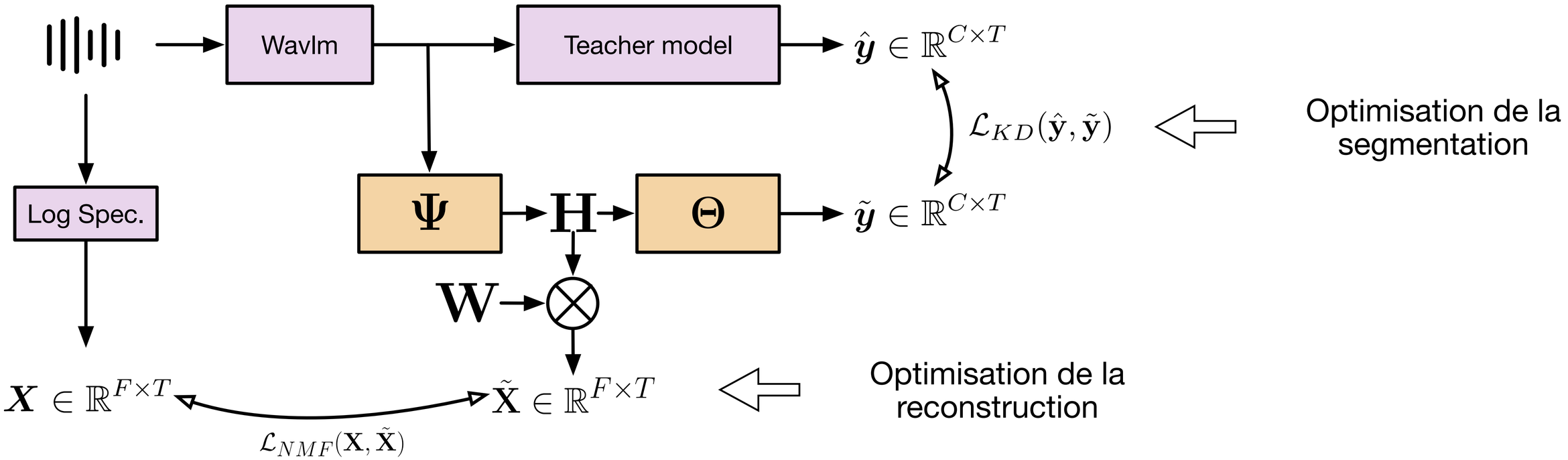
Distillation de connaissances



J. Parekh, et al. (2023) *Tackling Interpretability in Audio Classification Networks with Non-negative Matrix Factorization*. http://arxiv.org/abs/2305.07132

Segmentation audio explicable

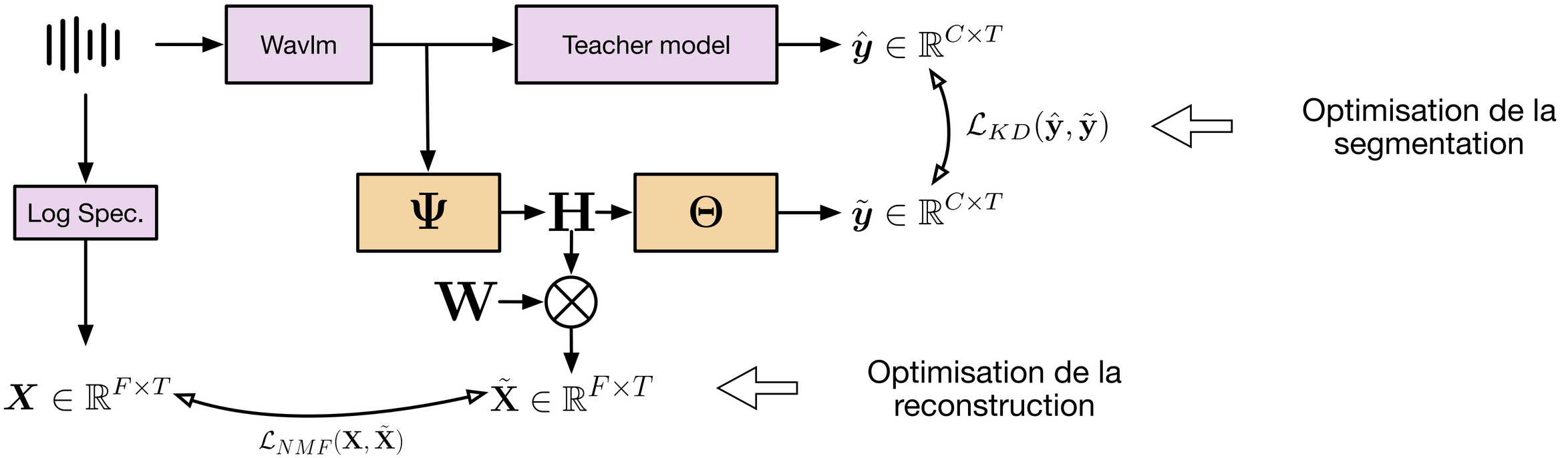
Distillation de connaissances



J. Parekh, et al. (2023) *Tackling Interpretability in Audio Classification Networks with Non-negative Matrix Factorization*. http://arxiv.org/abs/2305.07132

Segmentation audio explicable

Distillation de connaissances



$$\mathscr{L} = \alpha \mathscr{L}_{KD} + \beta \mathscr{L}_{NMF} + \gamma \|\mathbf{H}\|_1$$

J. Parekh, et al. (2023) *Tackling Interpretability in Audio Classification Networks with Non-negative Matrix Factorization*. http://arxiv.org/abs/2305.07132

Sommaire

Segmentation audio explicable

Protocole d’évaluation

Jeux de données pour l’apprentissage de chaque système :

	Model		Available label			
Dataset	T	P	SAD	MD	ND	OSD
Albayzin	✓	✓	✓	✓	✓	
OpenBMAT	✓			✓		
ALLIES	✓		✓			✓
DiHard III	✓	✓	✓			✓

⇨ Annotations manquantes ignorées pour le calcul de la fonction de perte

⇨ Évaluation des performances sur 3 jeux de données : ALLIES-clean, DIHARD III et Aragon Radio

Segmentation audio explicable

Protocole d'évaluation

Jeux de données pour l'apprentissage de chaque système :

	Model		Available label			
Dataset	T	P	SAD	MD	ND	OSD
Albayzin	✓	✓	✓	✓	✓	
OpenBMAT	✓			✓		
ALLIES	✓		✓			✓
DiHard III	✓	✓	✓			✓

Teacher

⇨ Annotations manquantes ignorées pour le calcul de la fonction de perte

⇨ Évaluation des performances sur 3 jeux de données : ALLIES-clean, DIHARD III et Aragon Radio

Segmentation audio explicable

Protocole d'évaluation

Jeux de données pour l'apprentissage de chaque système :

	Model		Available label			
Dataset	T	P	SAD	MD	ND	OSD
Albayzin	✓	✓	✓	✓	✓	
OpenBMAT	✓			✓		
ALLIES	✓		✓			✓
DiHard III	✓	✓	✓			✓

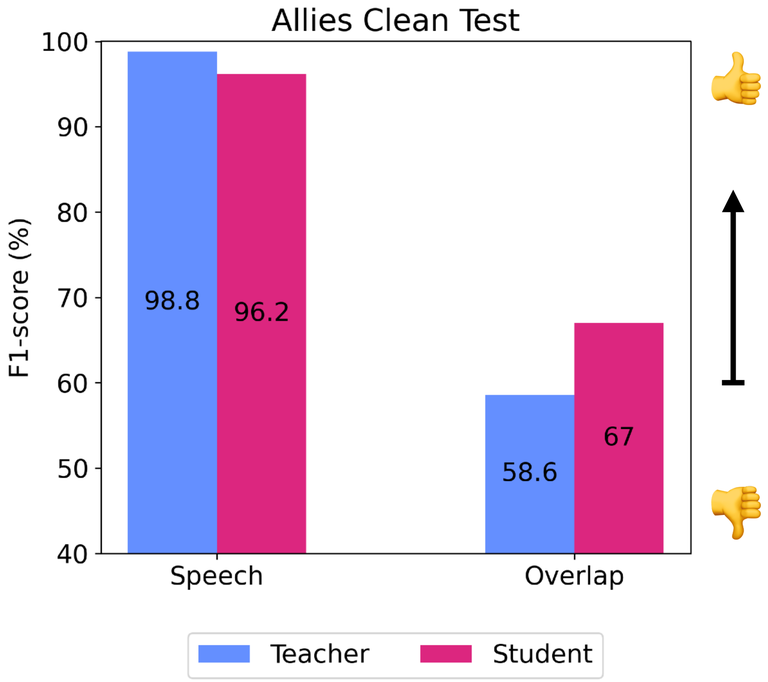
Teacher Student

⇨ Annotations manquantes ignorées pour le calcul de la fonction de perte

⇨ Évaluation des performances sur 3 jeux de données : ALLIES-clean, DIHARD III et Aragon Radio

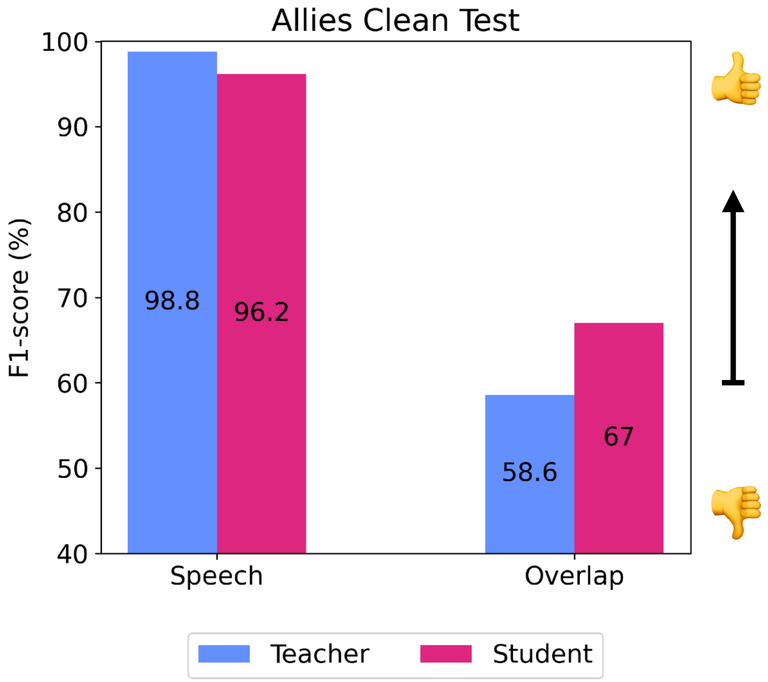
Performances de segmentation

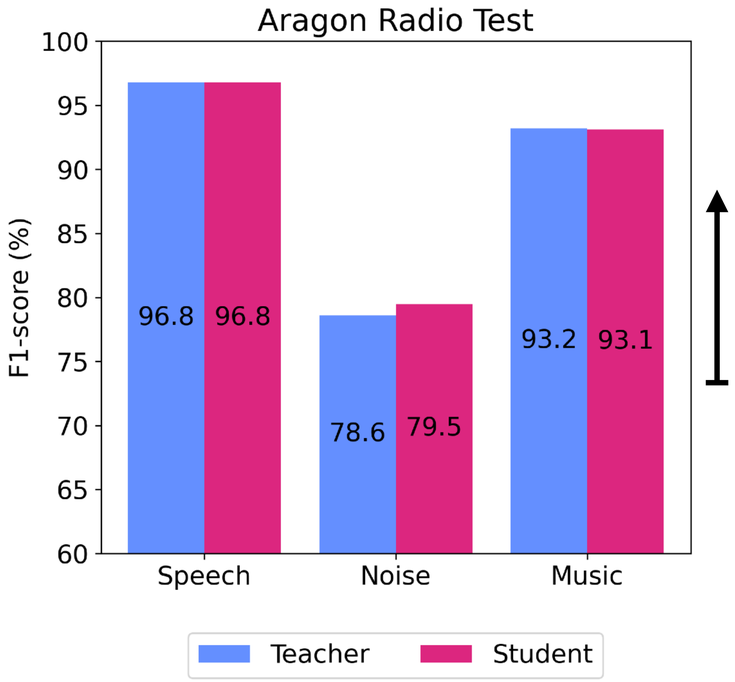
Teacher vs. *Student*



Performances de segmentation

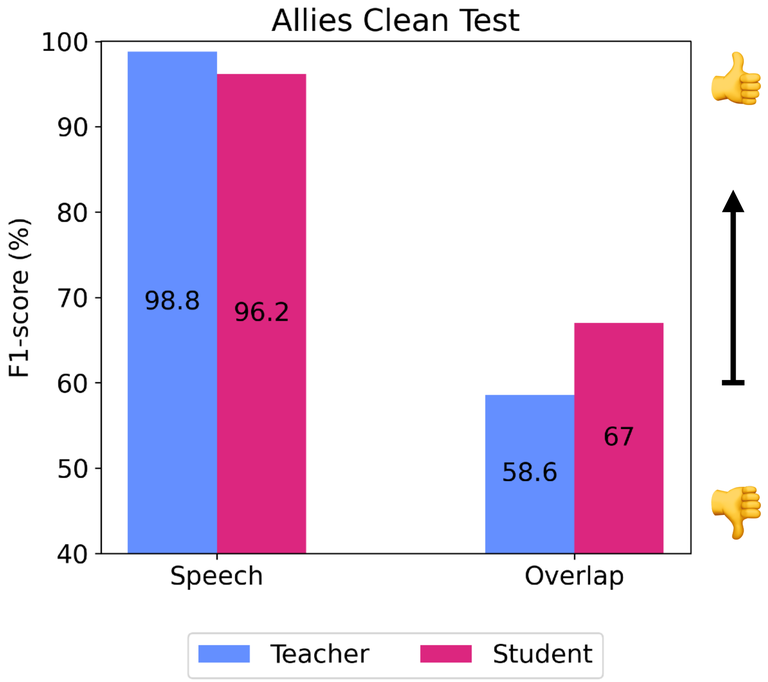
Teacher vs. *Student*

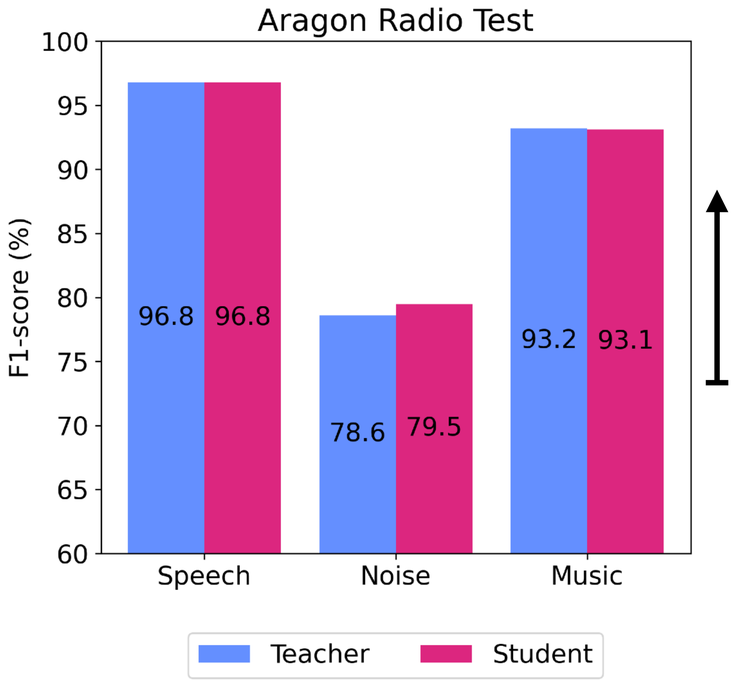


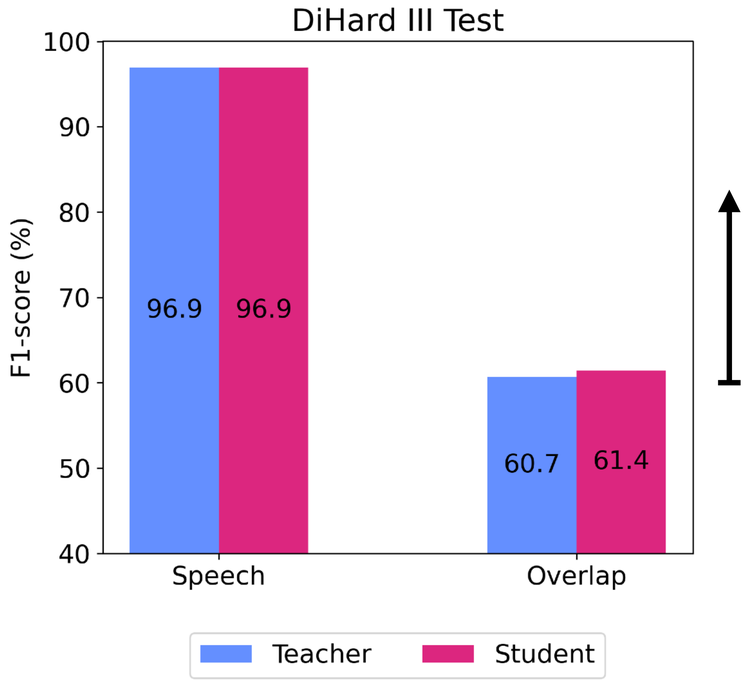


Performances de segmentation

Teacher vs. *Student*

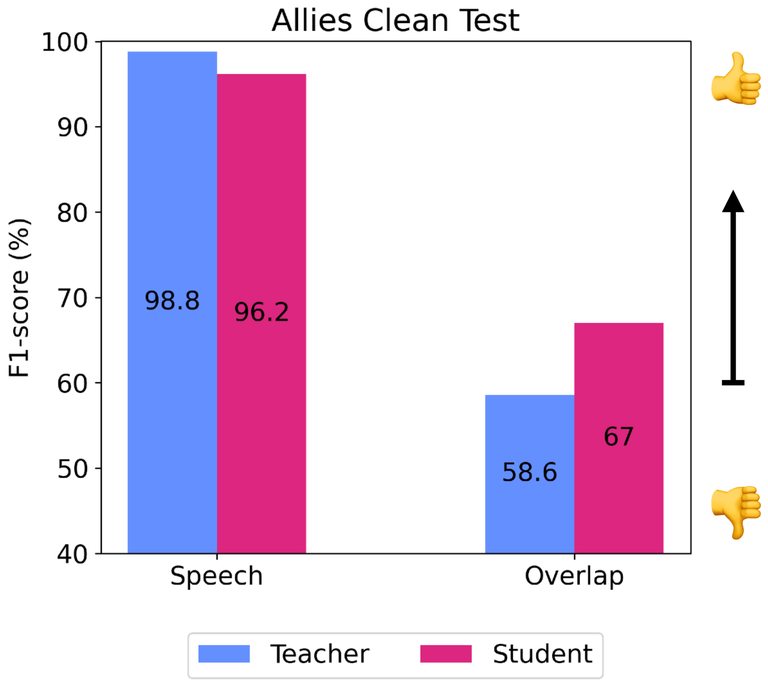


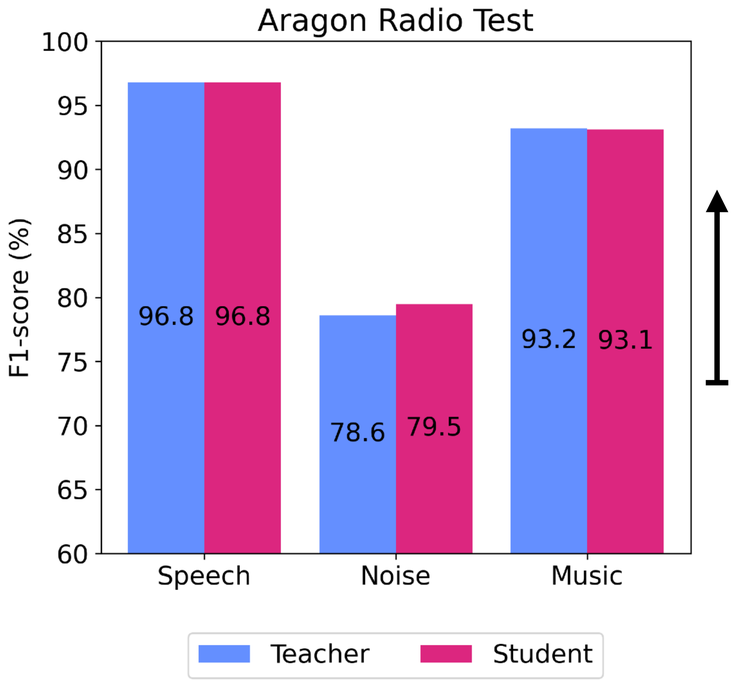


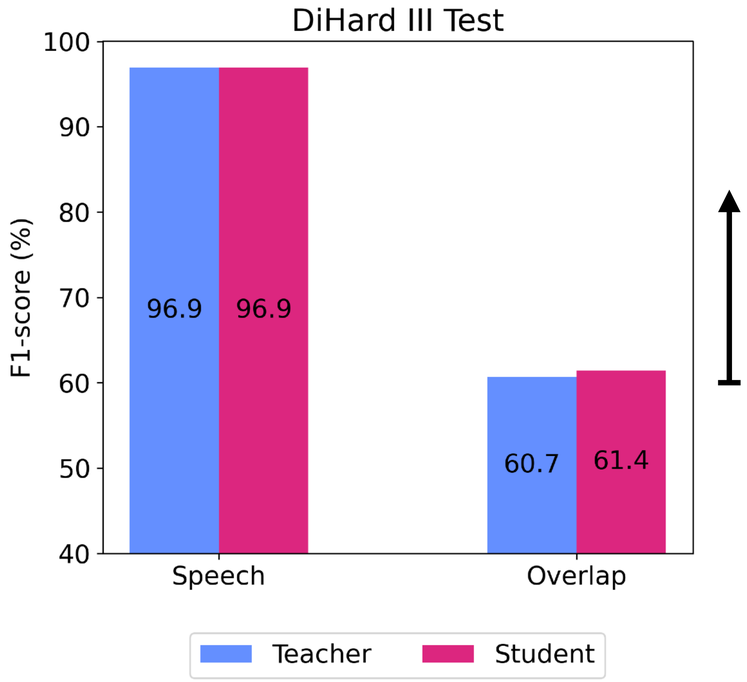


Performances de segmentation

Teacher vs. *Student*







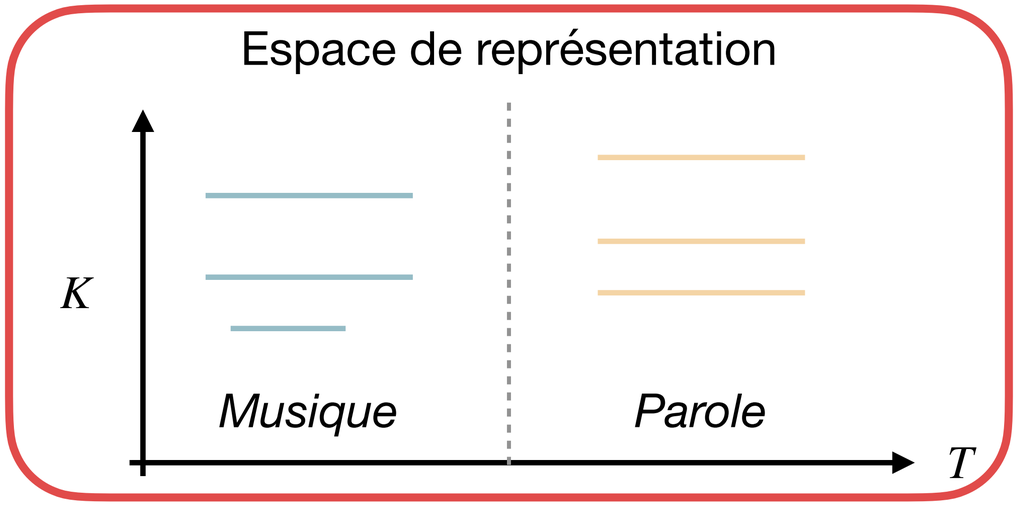
⇨ Le *student* offre les mêmes performances que le modèle original

Sommaire

Extraction d’explications

Passage de l’espace de représentation aux fréquences

$\hat{\mathbf{H}}$

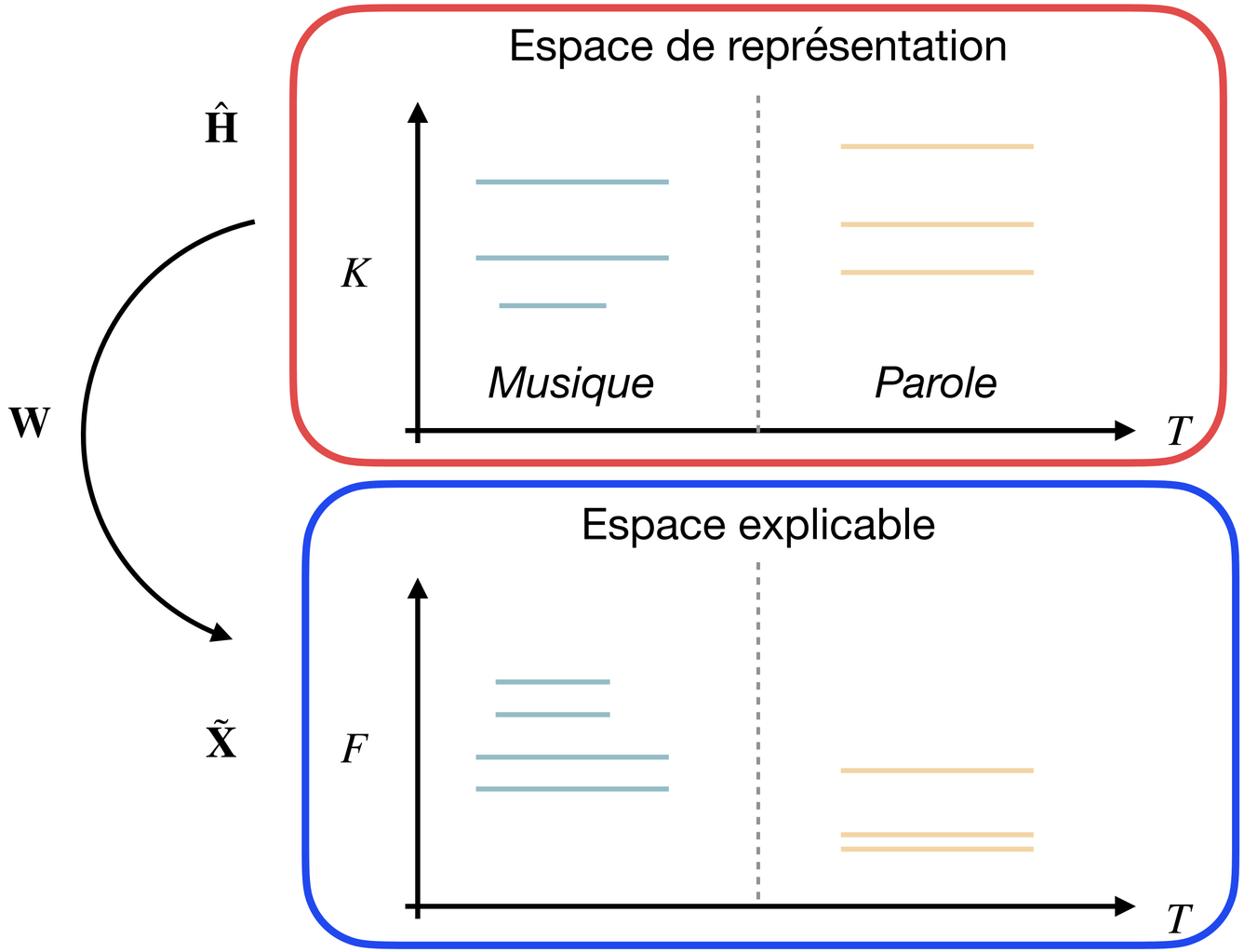


Sélection des composantes K pertinentes

$\tilde{\mathbf{X}}$

Extraction d’explications

Passage de l’espace de représentation aux fréquences

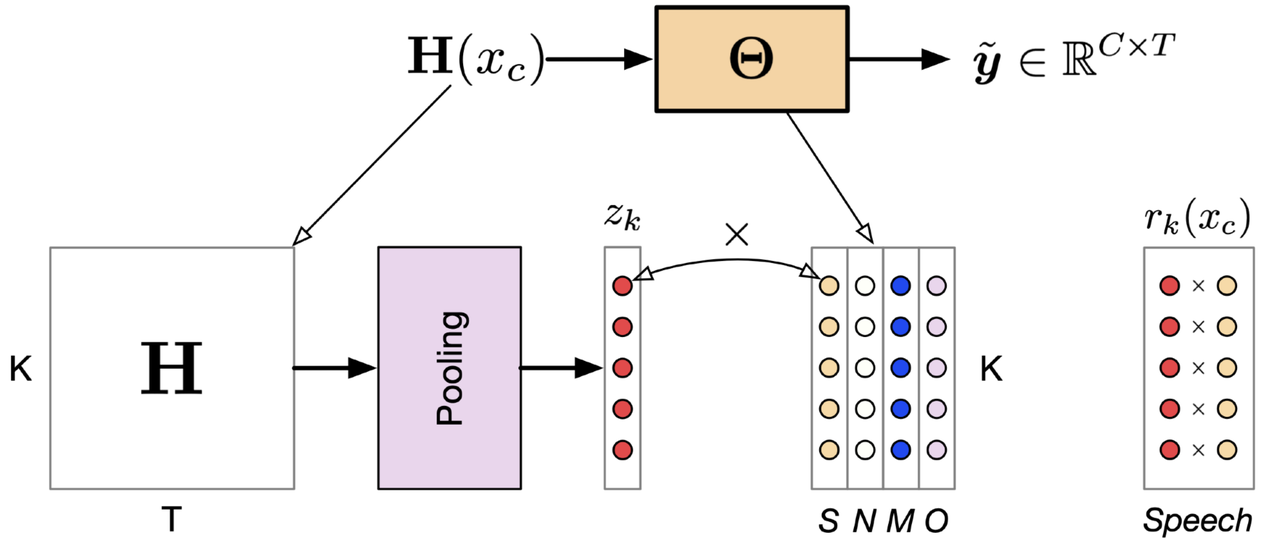


Sélection des composantes K pertinentes

Fréquences correspondantes

Extraction d’explications

Protocole

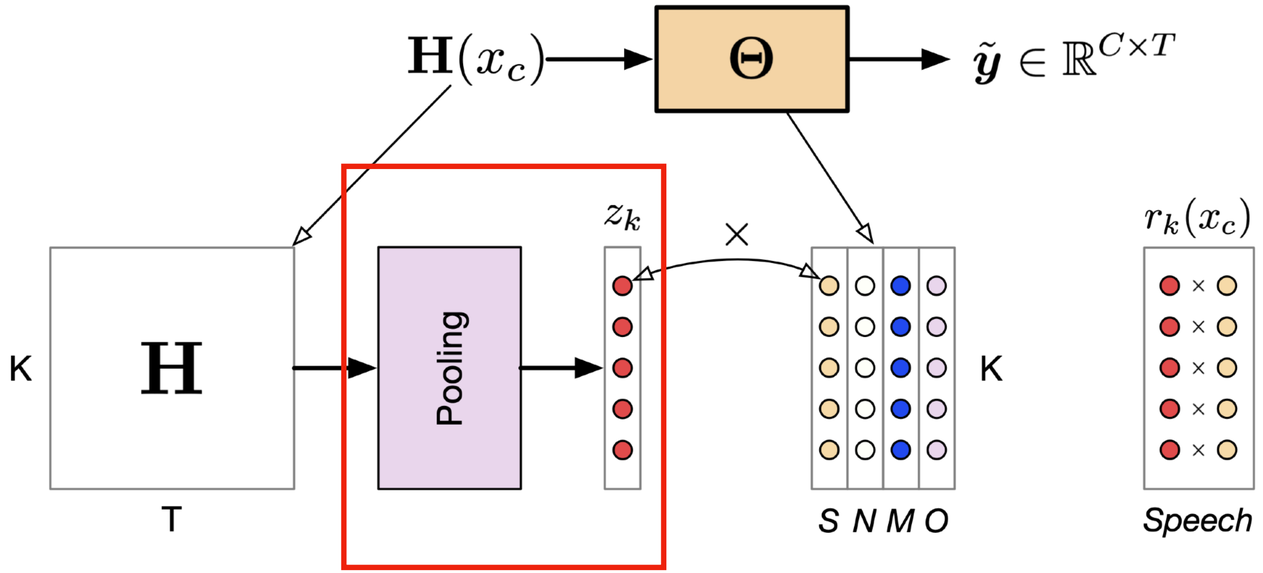


Extraction d’explications

Protocole

Moyenne temporelle de $\mathbf{H}$:

$$z_k\left(x_c\right) = \frac{1}{T}\sum_t H_{k,t}(x_c)$$

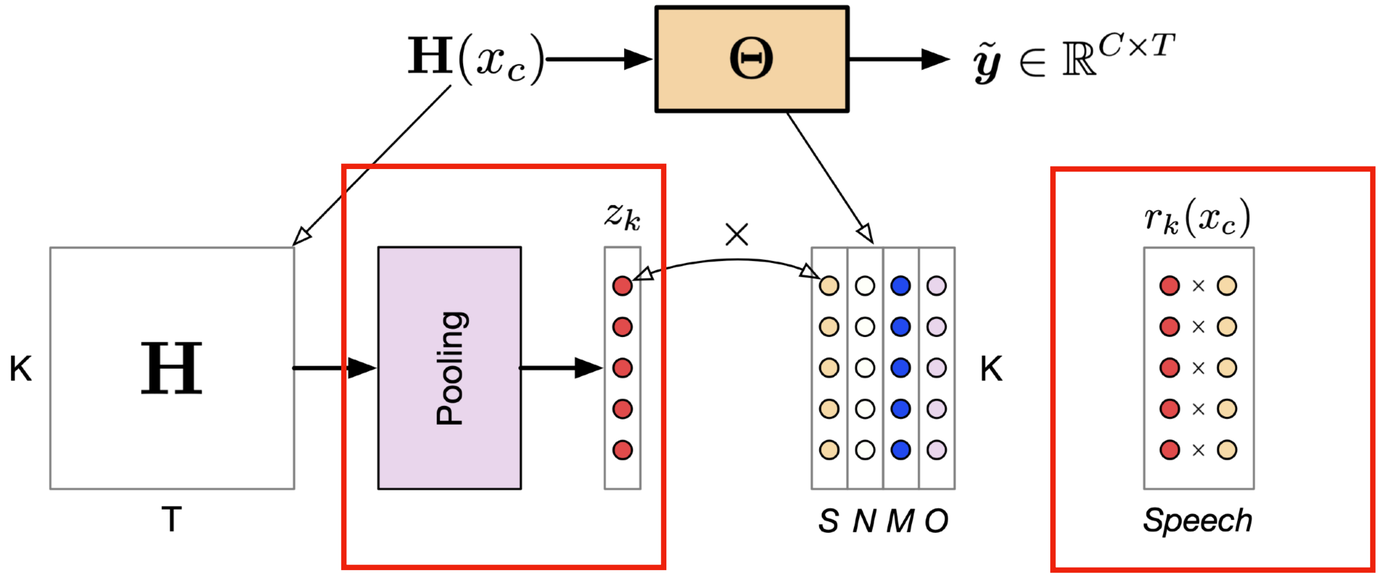


Extraction d’explications

Protocole

Moyenne temporelle de $\mathbf{H}$:

$$z_k\left(x_c\right) = \frac{1}{T}\sum_t H_{k,t}(x_c)$$



Mesure de la pertinence des composantes NMF pour détecter une classe c :

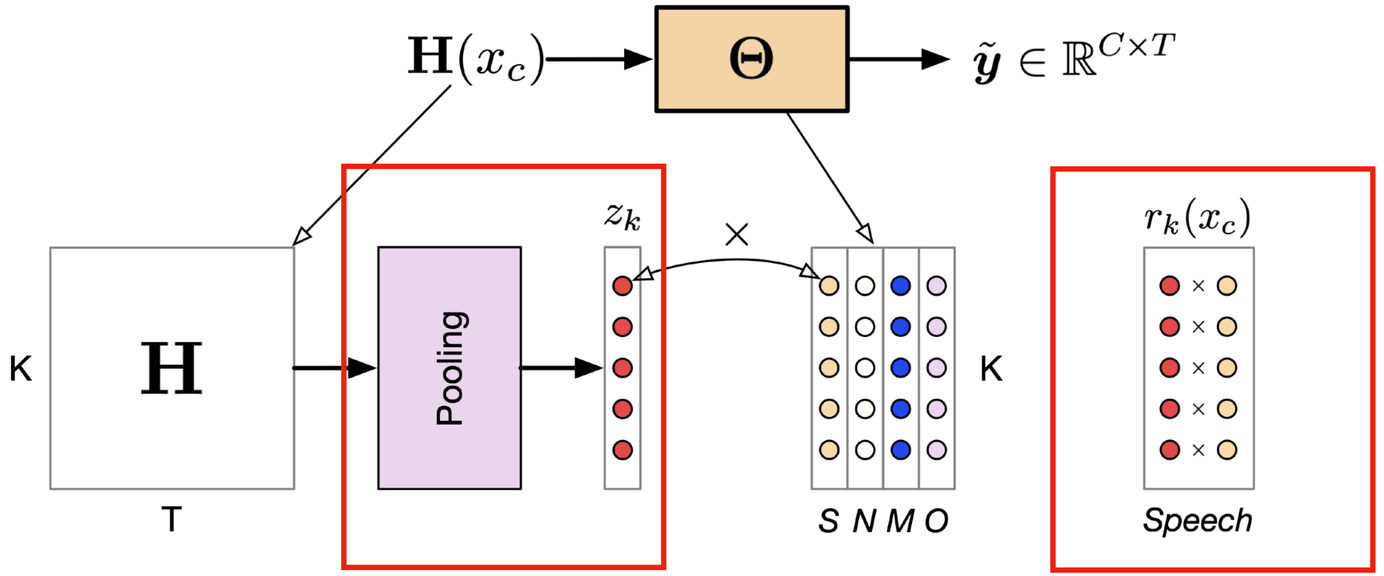
$$r_k\left(x_c\right) = z_k\left(x_c\right) \times \theta_{k,c}$$

Extraction d'explications

Protocole

Moyenne temporelle de $\mathbf{H}$:

$$z_k\left(x_c\right) = \frac{1}{T}\sum_t H_{k,t}(x_c)$$



Mesure de la pertinence des composantes NMF pour détecter une classe c :

$$r_k\left(x_c\right) = z_k\left(x_c\right) \times \theta_{k,c}$$

⇨ Un seuil peut être appliqué à r_k afin de sélectionner les composantes NMF

Extraction d'explications

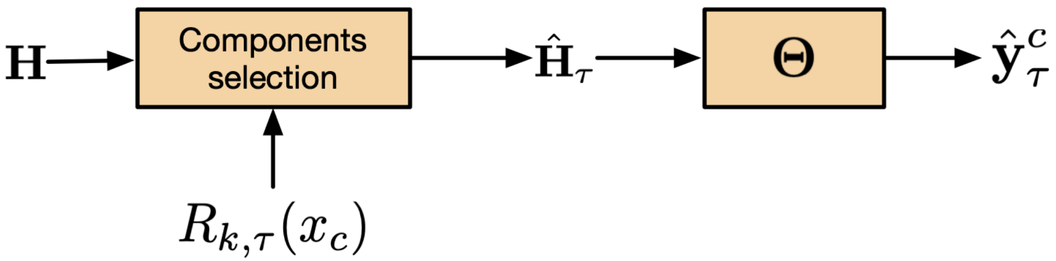
Protocole

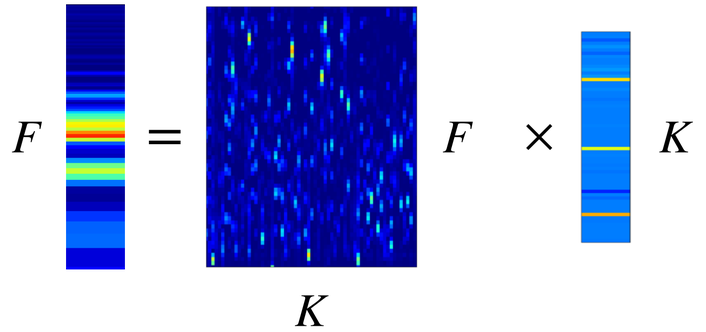
Rappel: $\mathbf{W}$ permet projeter les activations $\mathbf{H}$ dans le domaine des fréquences

$$\tilde{\mathbf{X}} = \mathbf{W}\mathbf{H}(x)$$

💡 Sélection des composantes pertinentes puis projection afin de le visualiser dans un espace connu

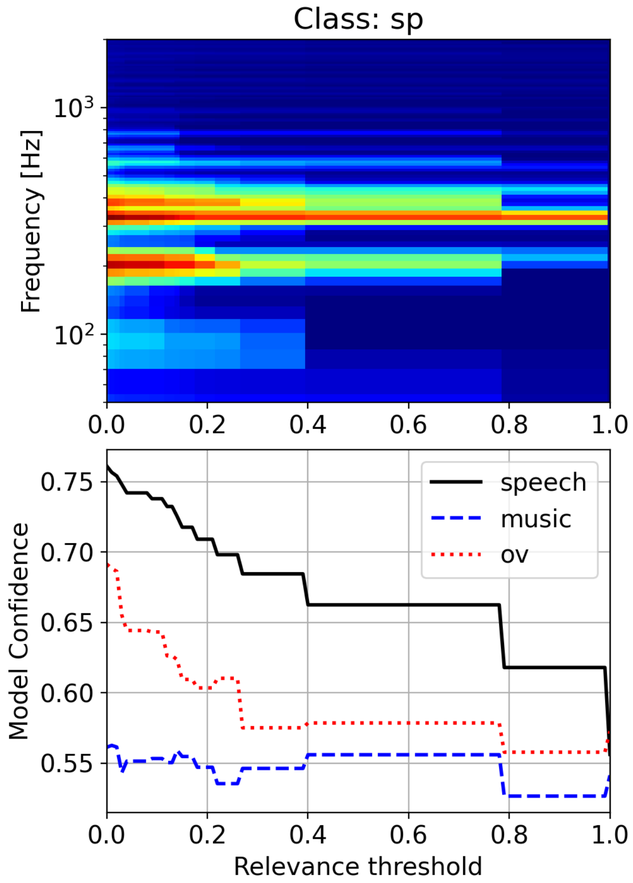
$$R_{k,c}(\tau) = \begin{cases} r_{k,c} & \text{if } r_{k,c} > \tau \\ 0 & \text{otherwise}. \end{cases} \quad \Rightarrow \quad \tilde{\mathbf{X}}_c = \mathbf{W}\mathbf{R}_\tau(x_c)$$





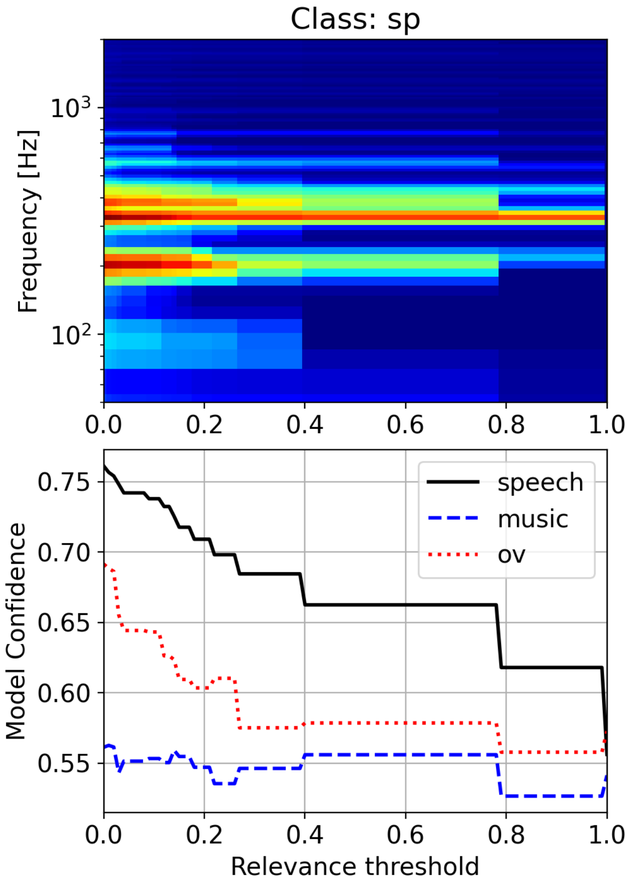
Extraction d’explications

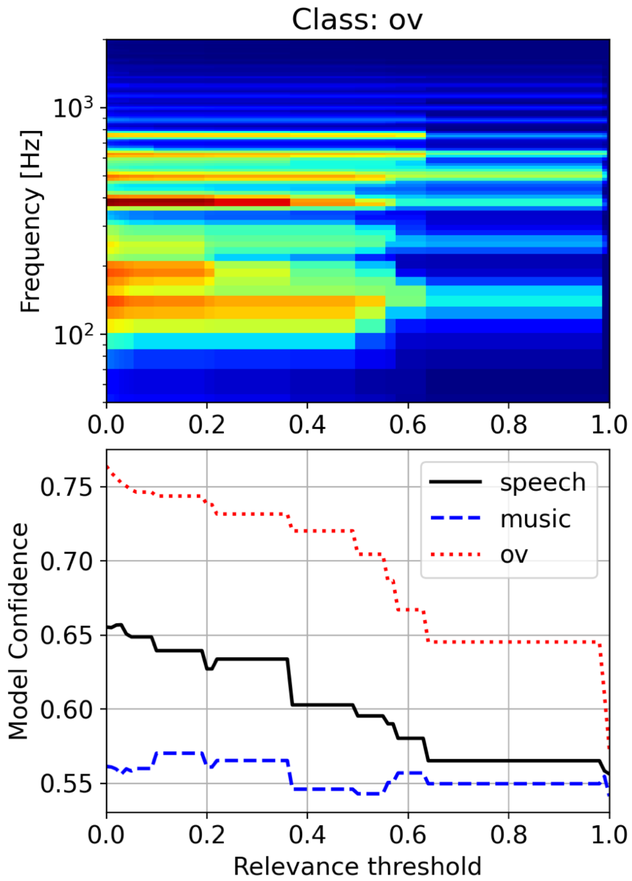
Explications locales : quelles fréquences sont utiles à la détection ?



Extraction d’explications

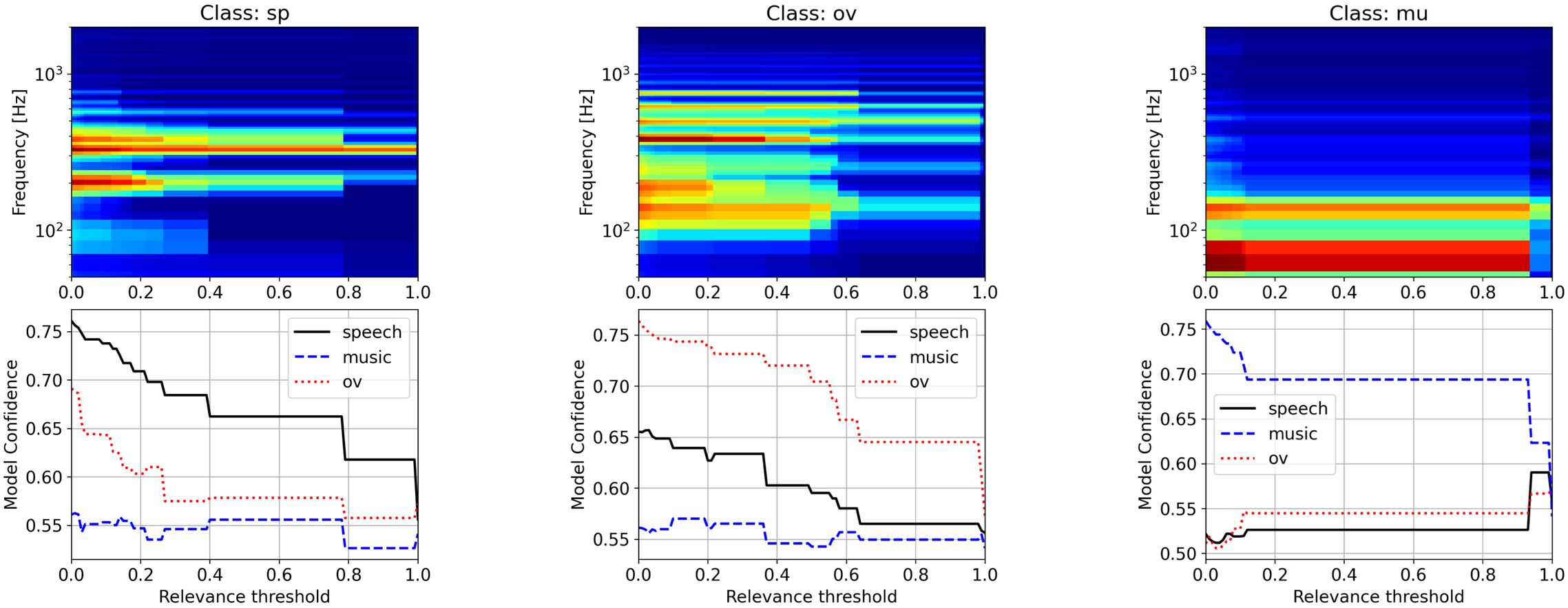
Explications locales : quelles fréquences sont utiles à la détection ?





Extraction d’explications

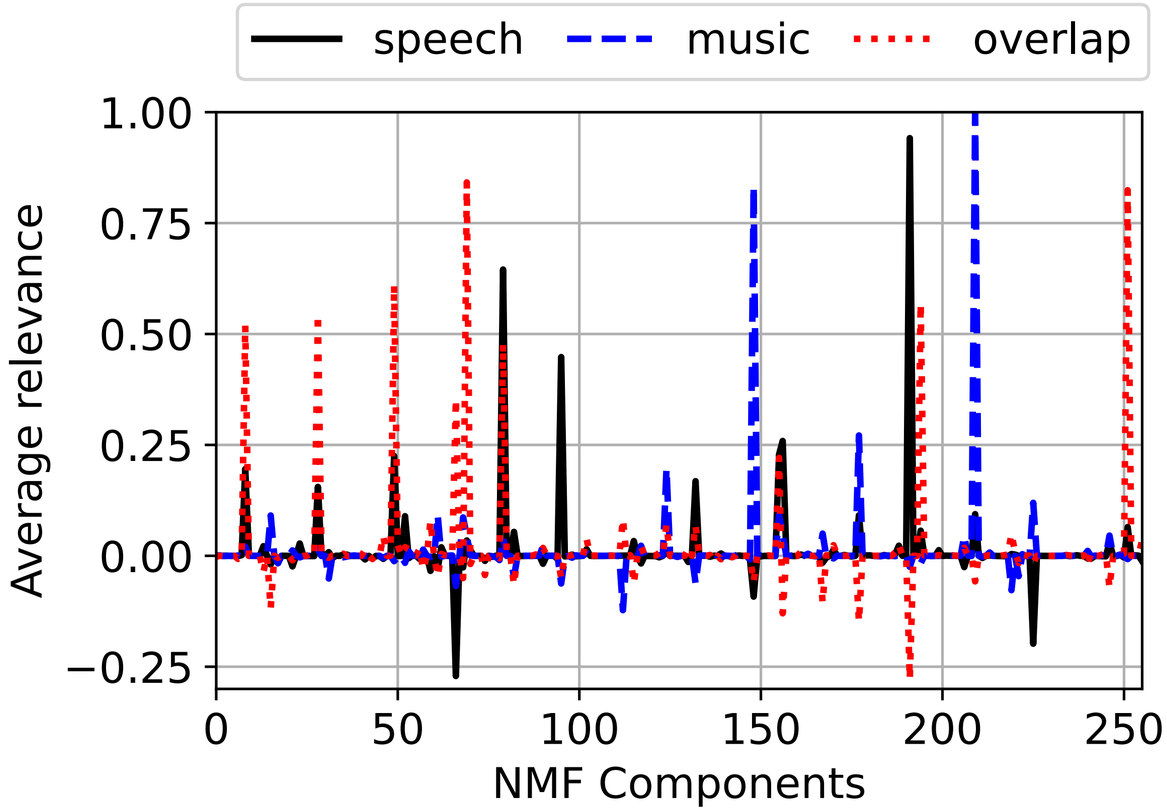
Explications locales : quelles fréquences sont utiles à la détection ?



Extraction d’explications

Explications globales : quel est l’ « ADN » des classes ?

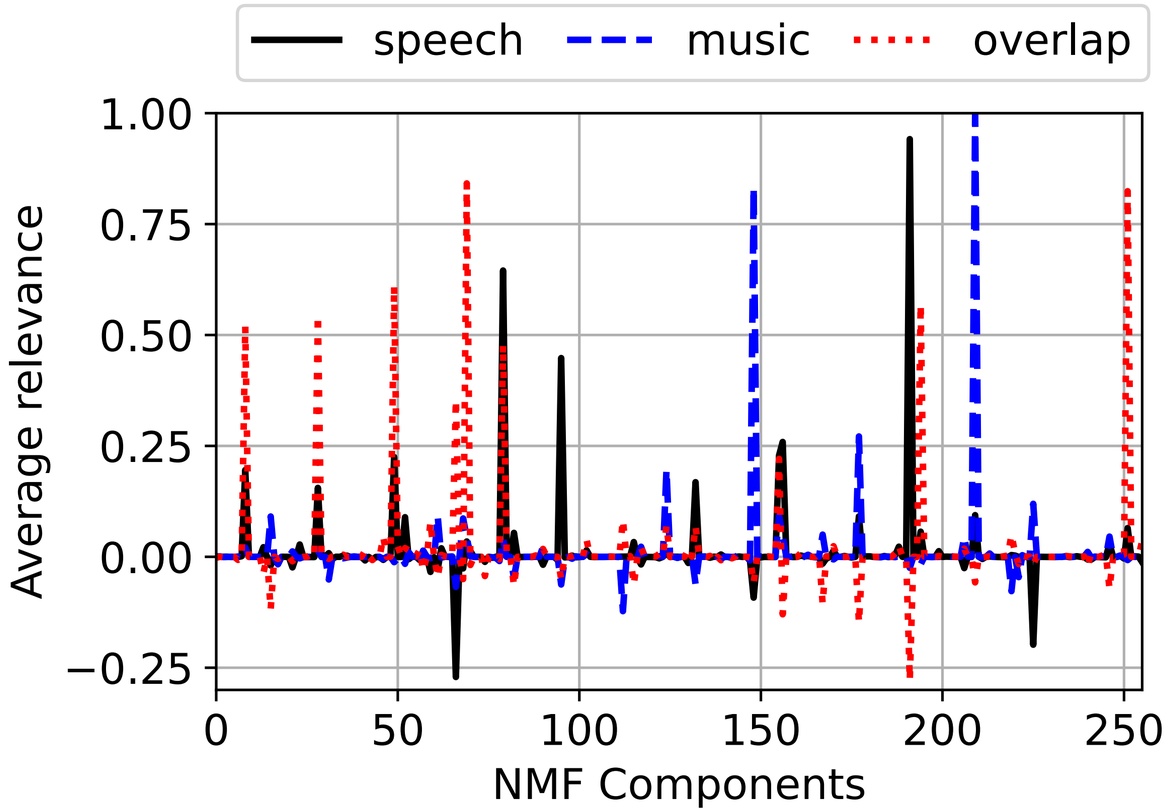
$$\mathbf{p}_c = \frac{1}{N_{\mathscr{D}}} \sum_{x_c \in \mathscr{D}} \mathbf{r}_c(x)$$



Extraction d’explications

Explications globales : quel est l’ « ADN » des classes ?

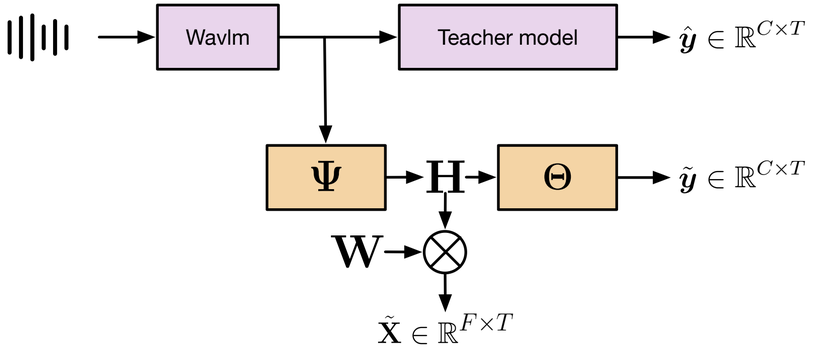
$$\mathbf{p}_c = \frac{1}{N_{\mathscr{D}}} \sum_{x_c \in \mathscr{D}} \mathbf{r}_c(x)$$

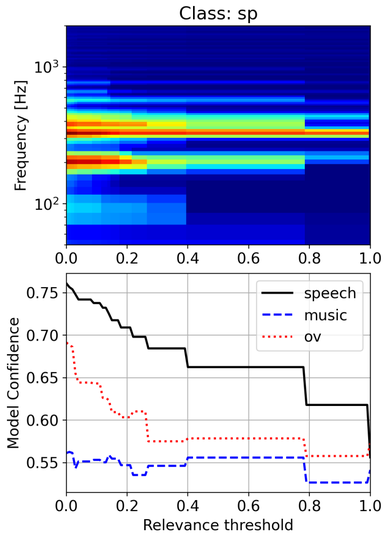


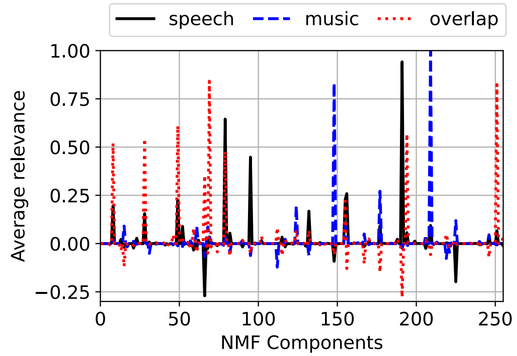
⇨ La pertinence moyenne pour chaque classe identifie les composantes représentatives

Conclusions

- Modèle de segmentation audio (parole, parole superposée, musique et bruit)
- Apprentissage d'un système explicable par distillation de connaissances
- Ce système utilise la NMF pour projeter l'espace de représentation dans le domaine des fréquences
- Mêmes performances que le *teacher*
- Possibilité d'extraire des explications locales (pour un exemple) et globales (pour un jeu de données)

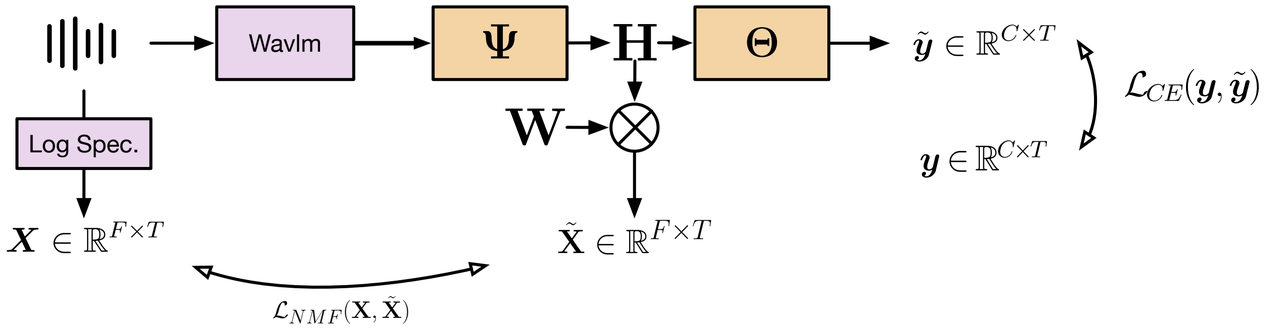






Perspectives

Supression du *teacher* pour définir un modèle explicable par construction



Probing de la matrice $\mathbf{H}$ pour une meilleure explication :
« cette matrice n'est-elle pas sur-optimisée ? »

Détections de genre de musique, de phonèmes, d'évènements sonores…