Supervised group nonnegative matrix factorisation with similarity constraints and applications to speaker identification

Romain Serizel[1], **Victor Bisot**[2], **Slim Essid**[2], **Gaël Richard**[2]

[1]LORIA, Univeristé de Lorraine, Inria, CNRS (France)
[2]LTCI, Télécom ParisTech, Université Paris-Saclay (France)

Monday 6th, March 2017

Outline

1. Speaker identification : What ? Why ?
2. Task-driven group NMF
3. Conclusions

Speaker identification

Main goal

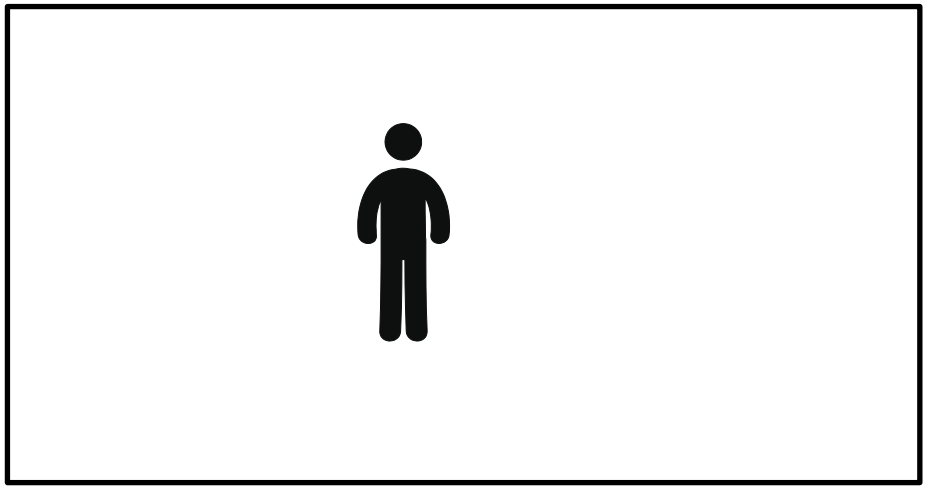
Identify a person from an audio recording

Speaker identification

Main goal

Identify a person from an audio recording

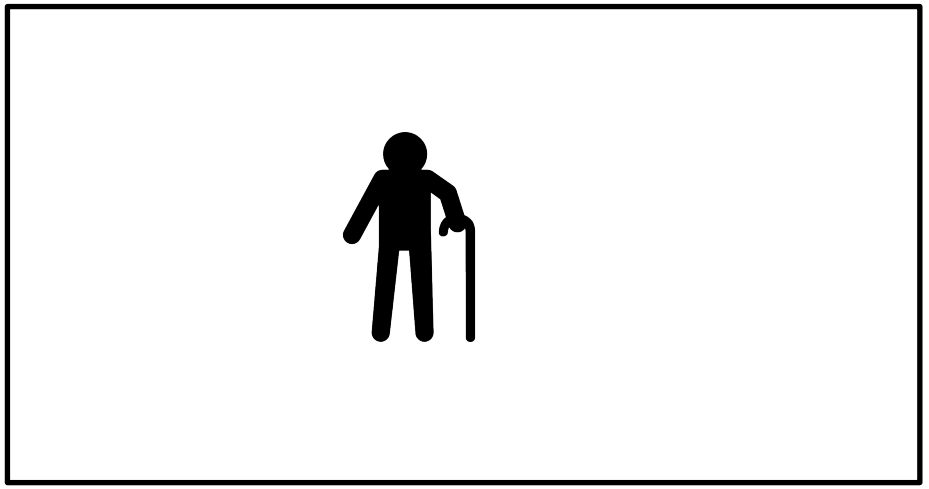
Session concept



Recording variability

Icons made by Freepik from www.flaticon.com

Session concept



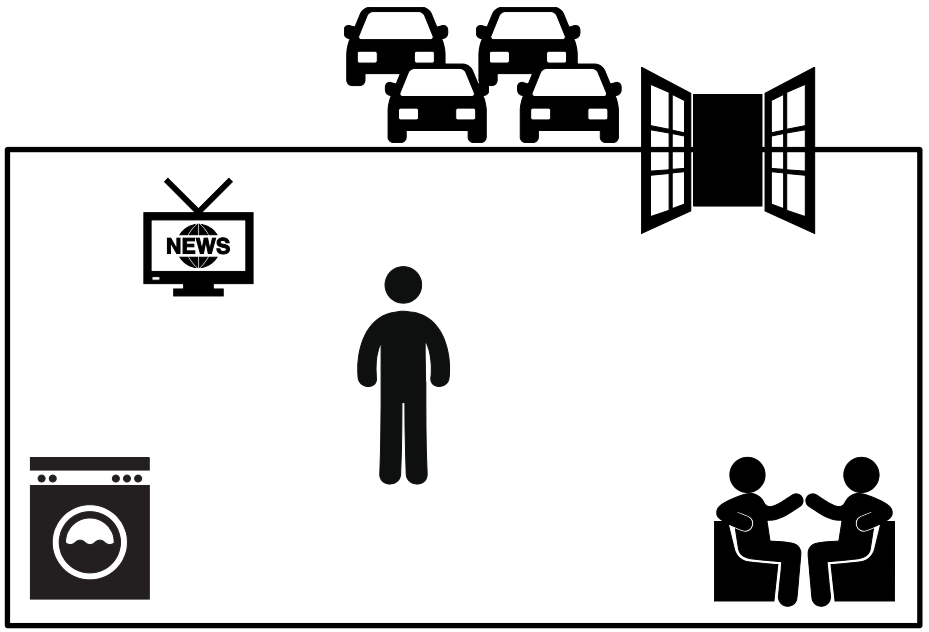
Recording variability

- Aging,

Loria

Icons made by Freepik from www.flaticon.com

Session concept



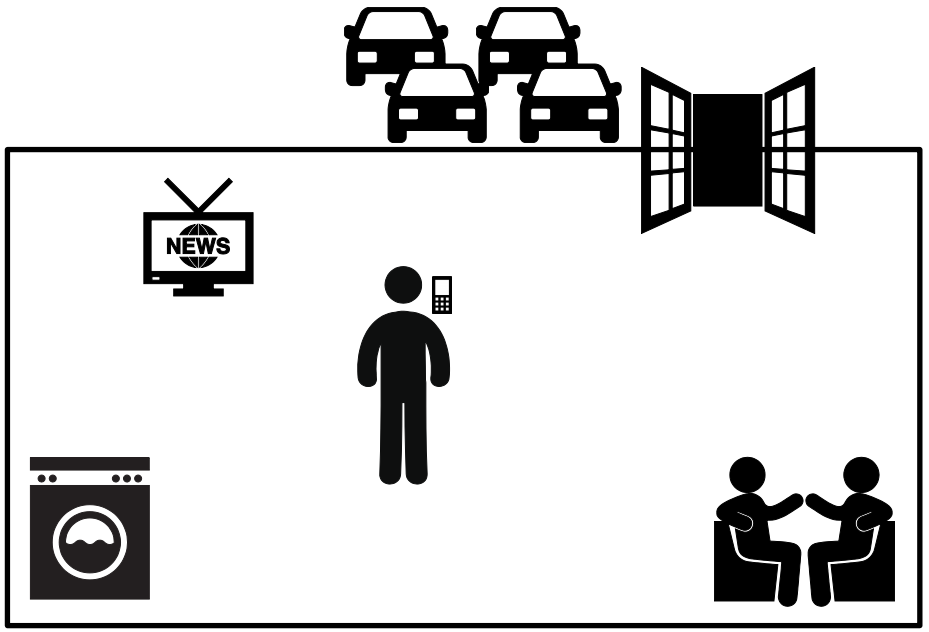
Recording variability

- Aging,
- Perturbations,

Loria

Icons made by Freepik from www.flaticon.com

Session concept



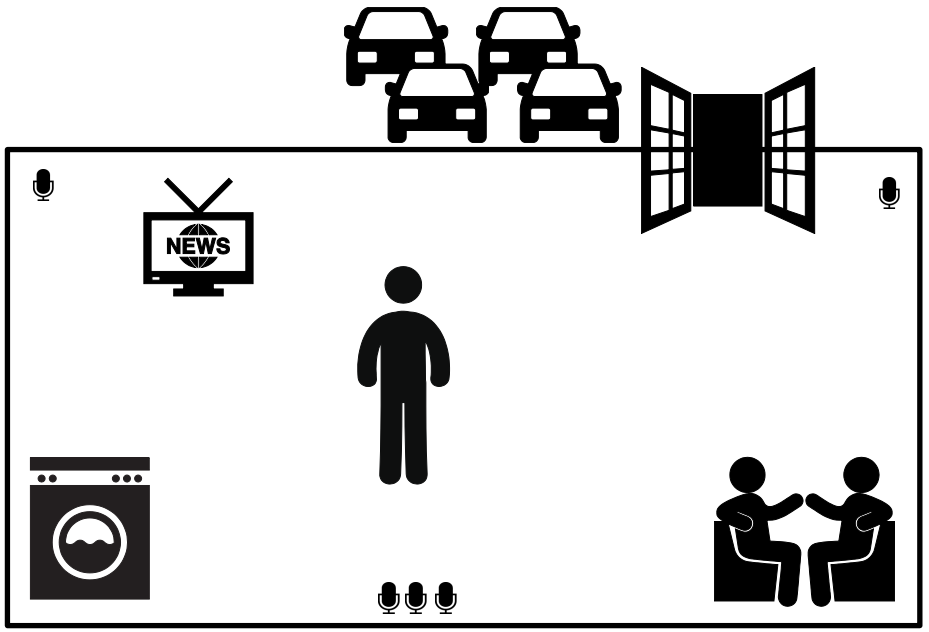
Recording variability

- Aging,
- Perturbations,
- Microphones. . .

Loria

Icons made by Freepik from www.flaticon.com

Session concept



Recording variability

- Aging,
- Perturbations,
- Microphones. . .

Icons made by Freepik from www.flaticon.com

Applications

Audio indexing (broadcast show, conferences,...)

- Content retrieval
- Rich-text transcription

Robust speech transcription

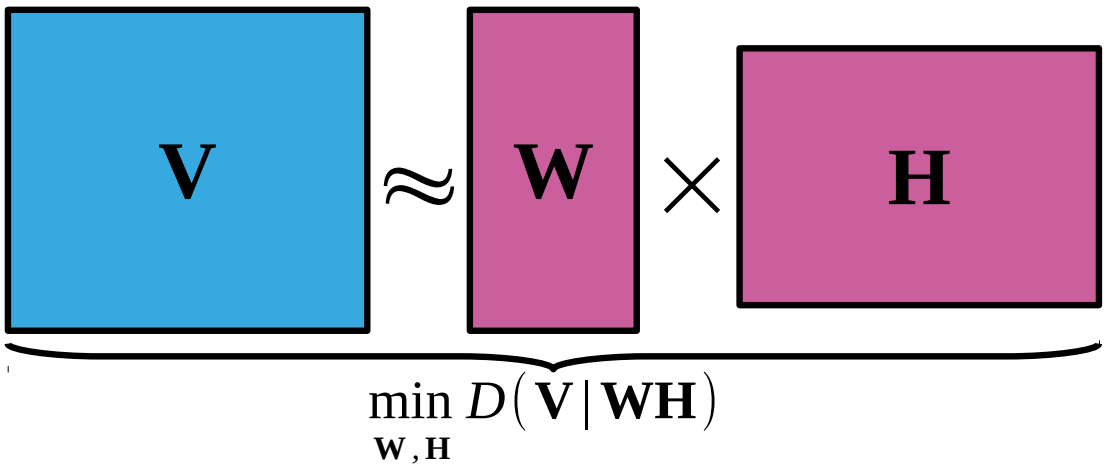
- Speaker adaptive training
- Speaker related feature/model adaptation

Voice-based identification

- Soft biometrics

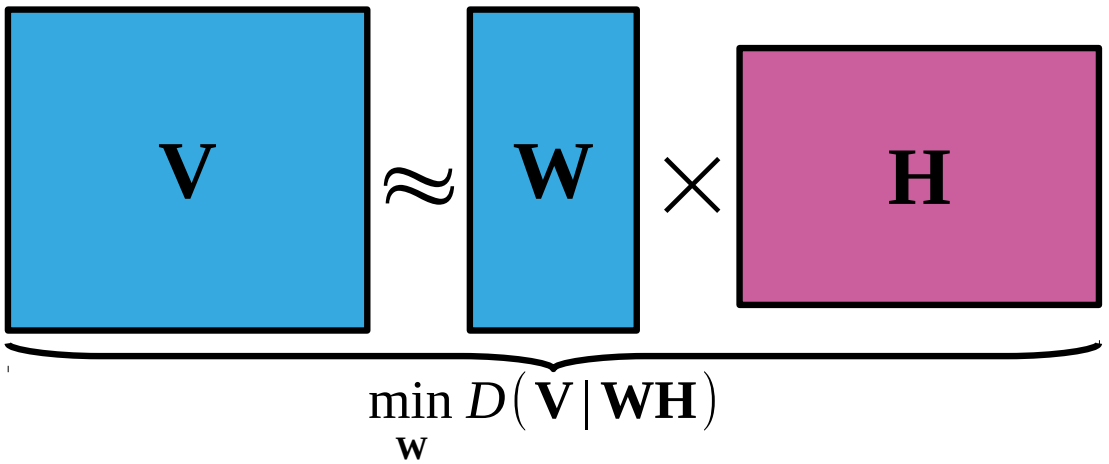
Standard classification chain (1)

- Train the dictionary



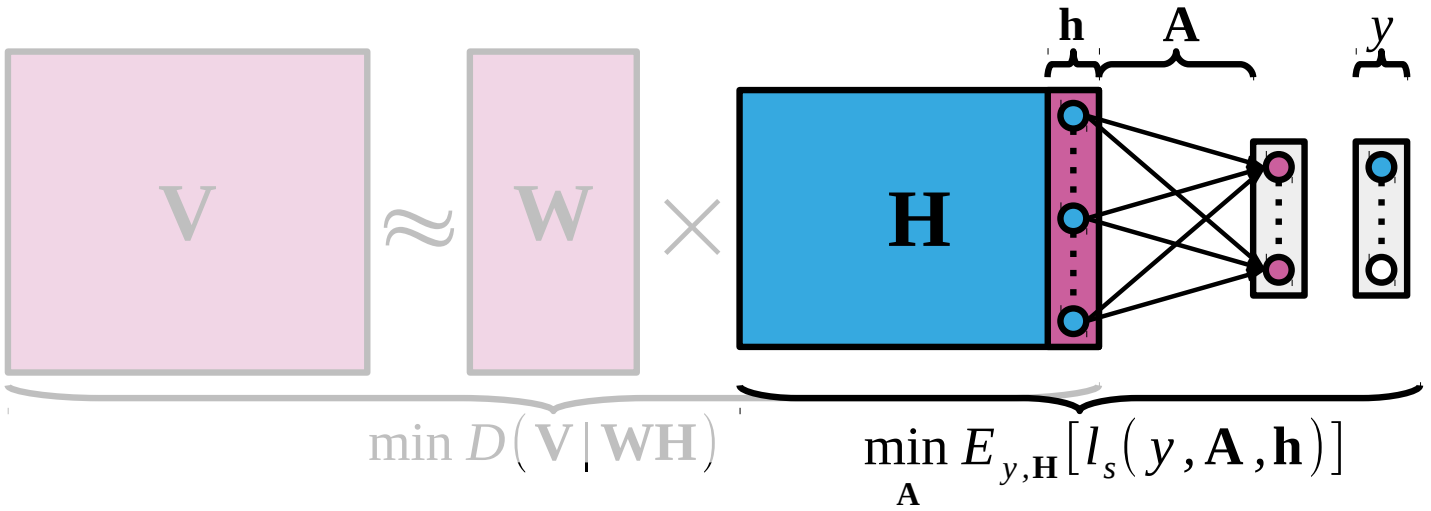
Standard classification chain (2)

- Project data on the new space



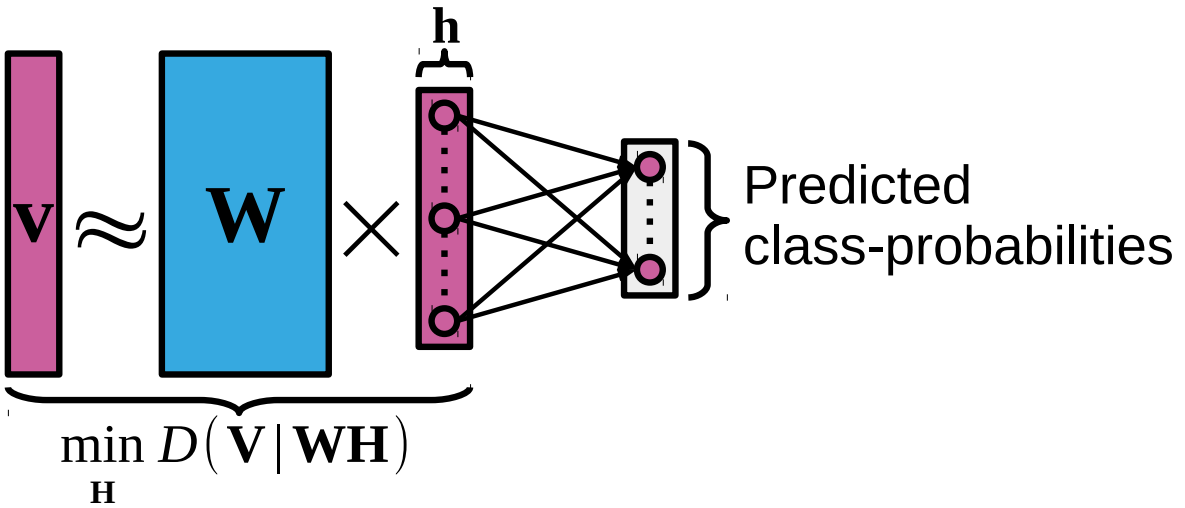
Standard classification chain (3)

- Train the classifier on projected data



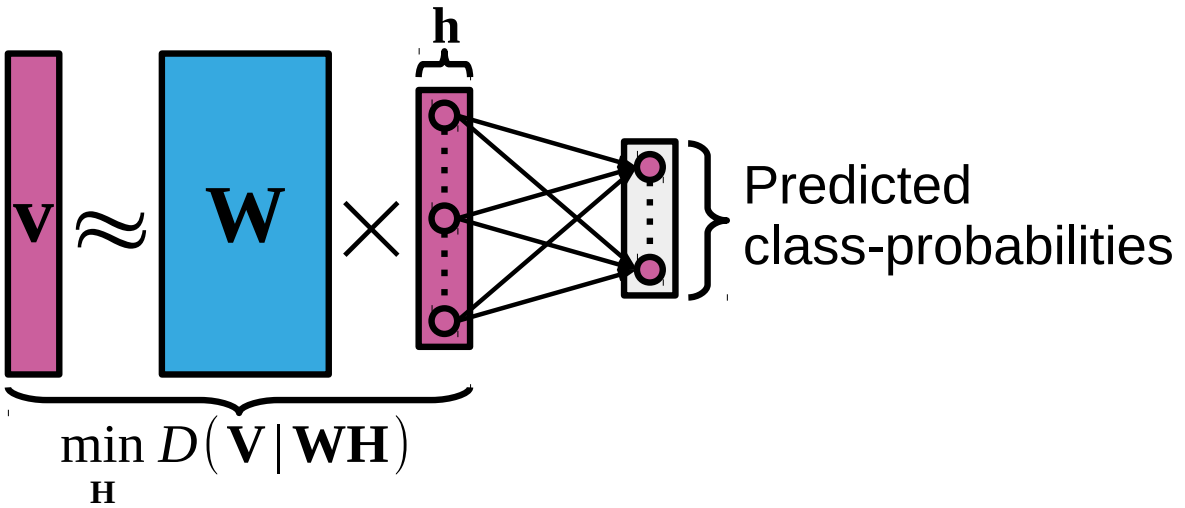
Standard classification chain (4)

- Use the dictionary and classifier



Standard classification chain (4)

- Use the dictionary and classifier



Problem

W and **H** are optimised according to a **reconstruction** criterion

Task-driven NMF (1)

General idea

Learn the dictionaries together with classifier parameters :

- Nested optimisation problem

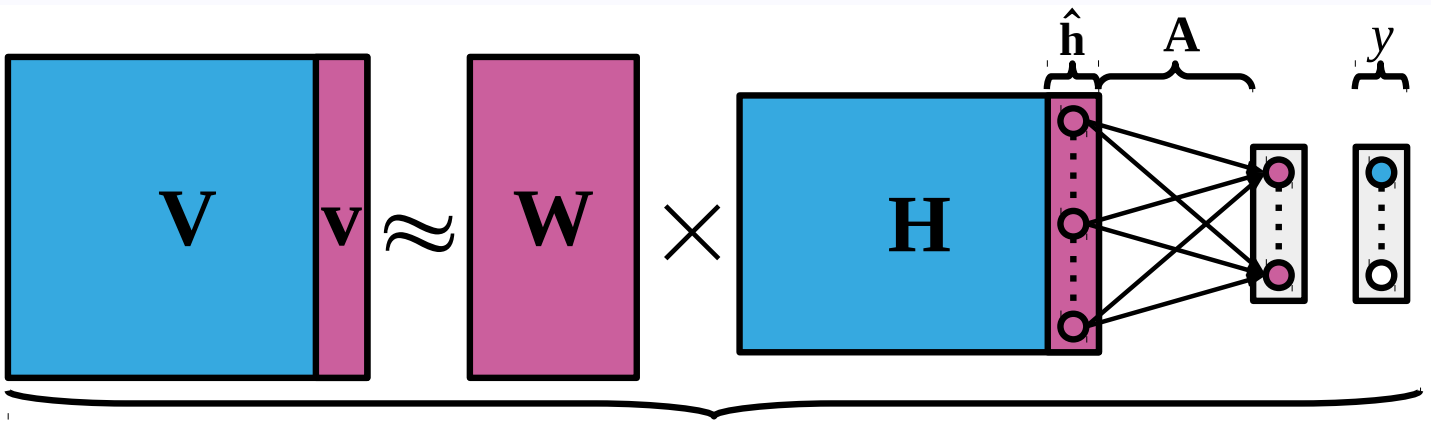
Dictionary divergence

- Euclidean norm : **closed form solution** for dictionaries
  - Task driven dictionary learning (Mairal et al., 2012)
  - Application to audio scene analysis (Bisot et al., 2016)
- General β-divergence :
  - Application to source separation (Sprechmann et al., 2014)
  - Application to event detection (Bisot et al., 2017)

Loria

Task-driven NMF (2)

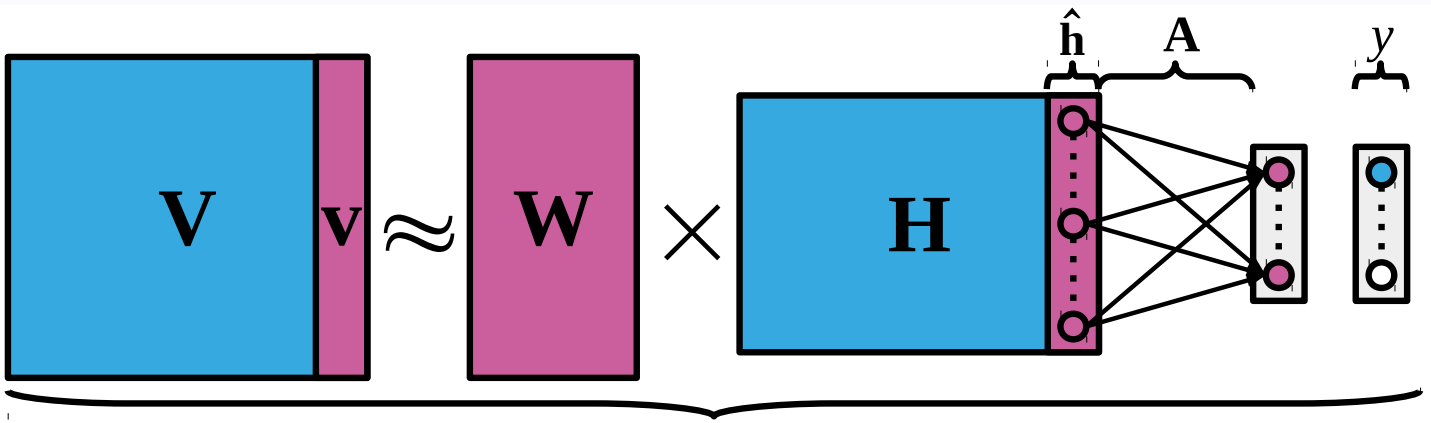
Task-driven NMF : General idea



$$\min_{\mathbf{W},\mathbf{A}} E_{y,\mathbf{v}}[l_s(y,\mathbf{A},\hat{\mathbf{h}}(\mathbf{v},\mathbf{W}))]$$

Task-driven NMF (2)

Task-driven NMF : General idea

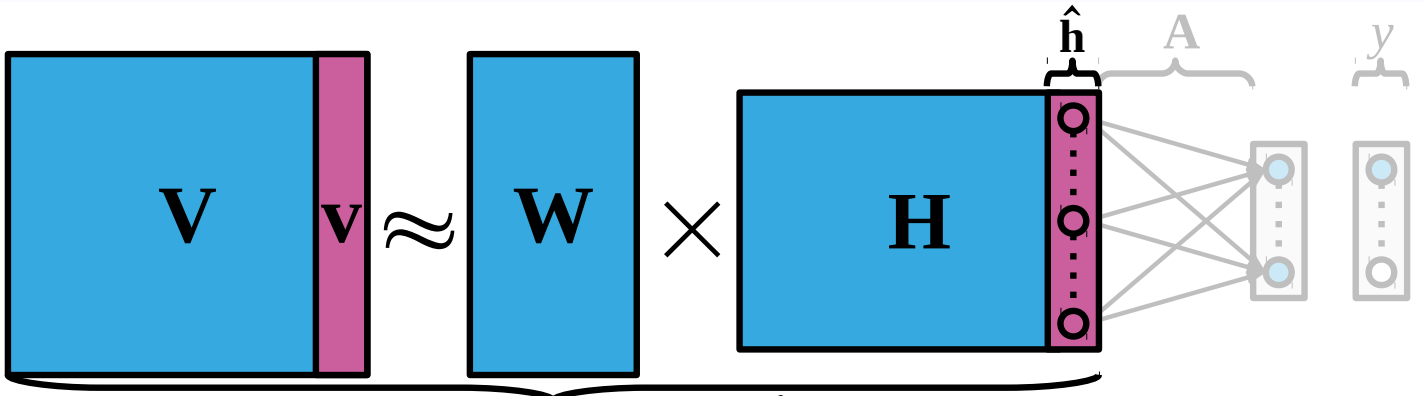


$$\min_{\mathbf{W},\mathbf{A}} E_{y,\mathbf{v}}[l_s(y,\mathbf{A},\hat{\mathbf{h}}(\mathbf{v},\mathbf{W}))]+\frac{\nu}{2}\|\mathbf{A}\|_2^2$$

Task-driven NMF : algorithm (1)

- For each new sample (**v**)

Project v on $\hat{h}$

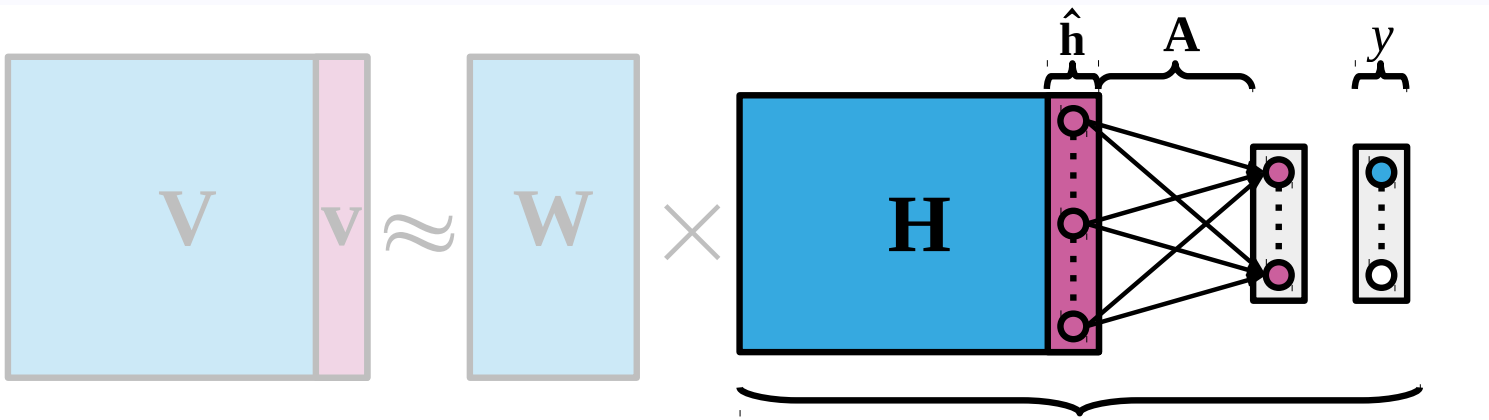


$$\min_{\mathbf{h}} \frac{1}{2}\|\mathbf{v}-\mathbf{W}\mathbf{h}\|_2^2+\lambda_1\|\mathbf{h}\|_1+\frac{\lambda_2}{2}\|\mathbf{h}\|_2^2$$

Task-driven NMF : algorithm (2)

- For each new sample (**v**)

Update the classifier parameters A

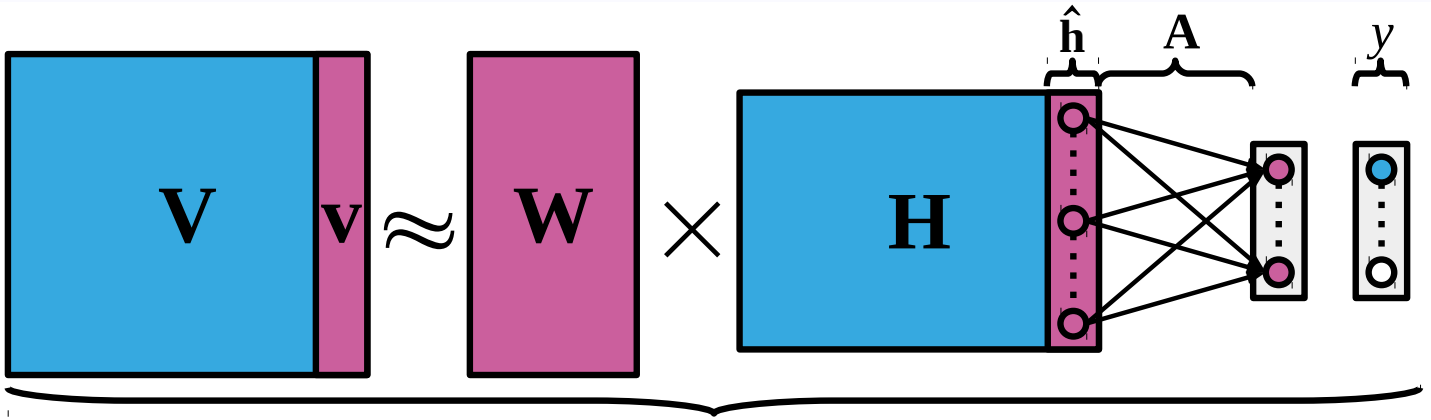


$$\min_{\mathbf{A}} E_{y,\mathbf{v}}[l_s(y,\mathbf{A},\hat{\mathbf{h}}(\mathbf{v},\mathbf{W}))]+\frac{\nu}{2}\|\mathbf{A}\|_2^2$$

Task-driven NMF : algorithm (3)

- For each new sample ($\mathbf{v}$)

Update the dictionary W



$$\min_{\mathbf{W}} E_{y,\mathbf{v}}[l_s(y,\mathbf{A},\hat{\mathbf{h}}(\mathbf{v},\mathbf{W}))]+\frac{\nu}{2}\|\mathbf{A}\|_2^2$$

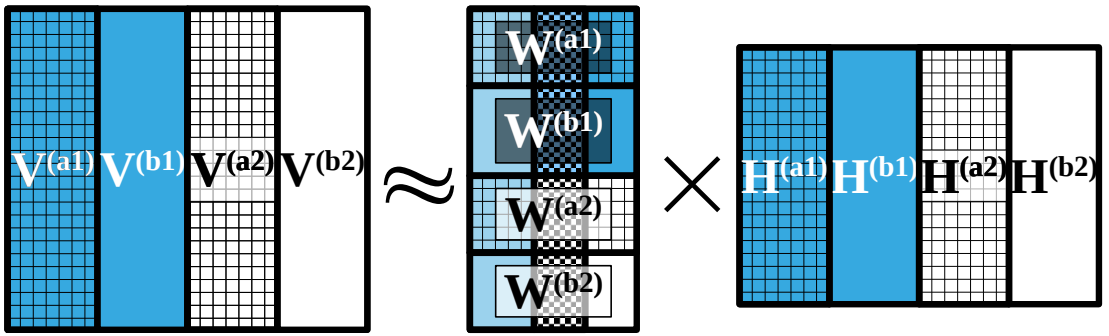
Task-driven NMF in practice

Implementation details

- Can be applied to sample or mini-batch
- Supports nonnegativity constraints for **W** and **H**
- Dictionary (**W**) initialisation :
  - Random
  - NMF
  - Concatenated group NMF dictionaries (Serizel et al., 2016)

Task-driven group NMF (1)

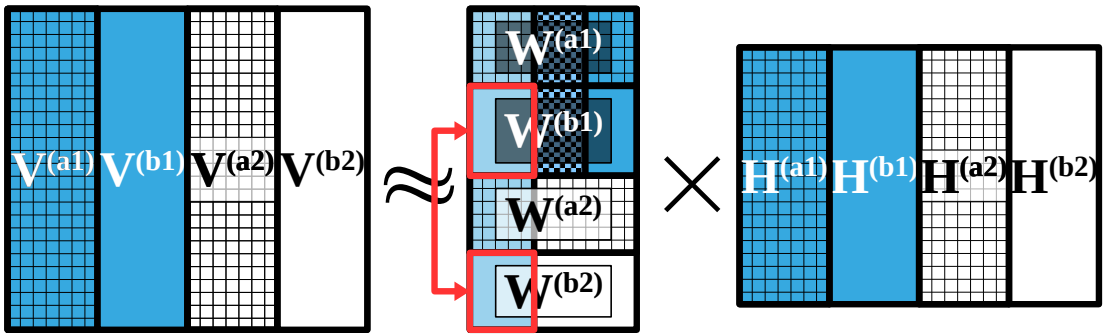
Include group NMF in a task-driven framework (Serizel et al., 2016)



- Subdictionnaries (related to a single speaker/session)

Task-driven group NMF (1)

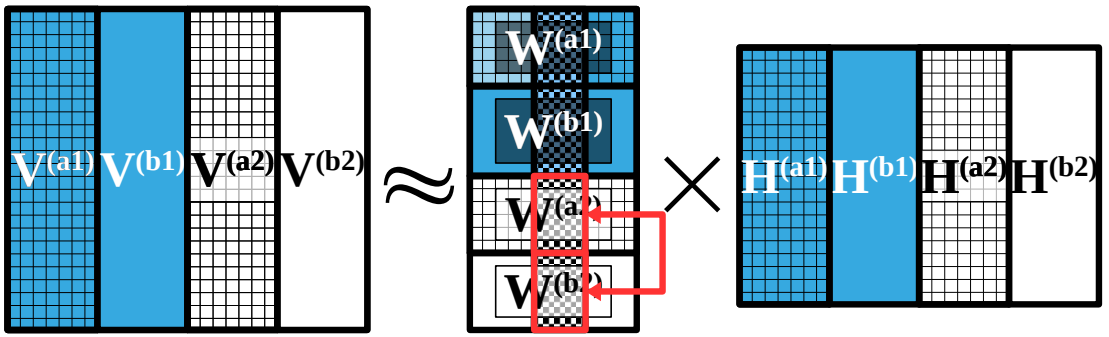
Include group NMF in a task-driven framework (Serizel et al., 2016)



- Subdictionnaries (related to a single speaker/session)
- Impose speaker/session similarity constraints

Task-driven group NMF (1)

Include group NMF in a task-driven framework (Serizel et al., 2016)

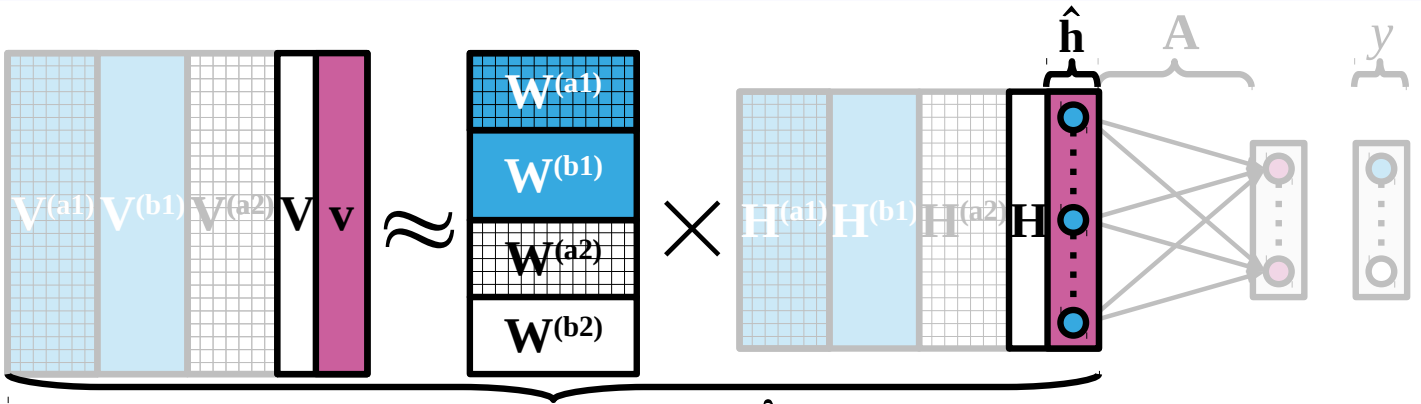


- Subdictionnaries (related to a single speaker/session)
- Impose speaker/session similarity constraints

Task-driven group NMF : algorithm (1)

- For each new sample ($\mathbf{v}$)

Project v on $\hat{\mathbf{h}}$



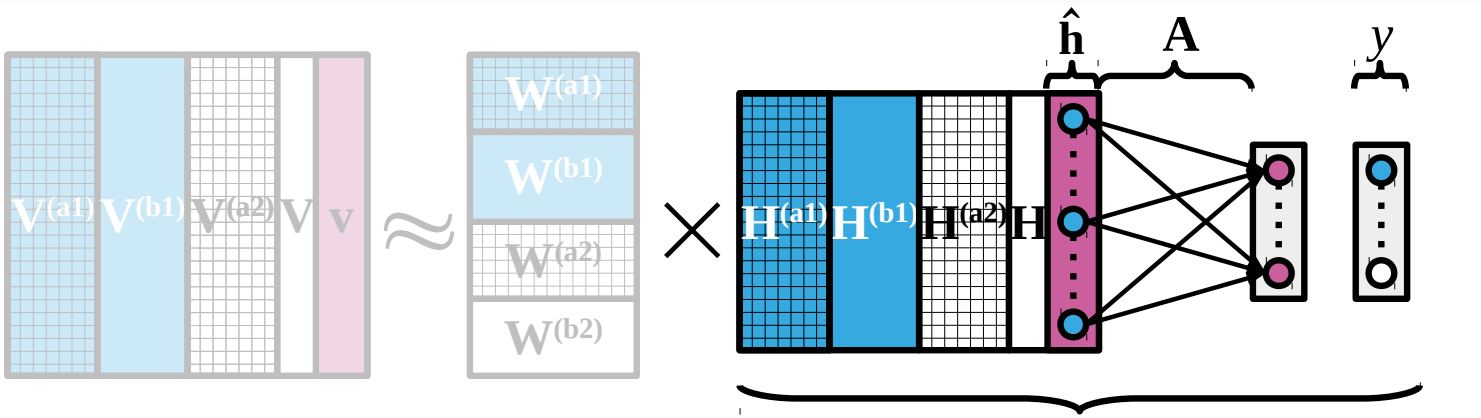
$$\min_{\mathbf{h}} \frac{1}{2}\|\mathbf{v}-\mathbf{W}\mathbf{h}\|_2^2+\lambda_1\|\mathbf{h}\|_1+\frac{\lambda_2}{2}\|\mathbf{h}\|_2^2$$

Source code is available at `https://github.com/rserizel/TGNMF`

Task-driven group NMF : algorithm (2)

- For each new sample (**v**)

Update the classifier parameters A



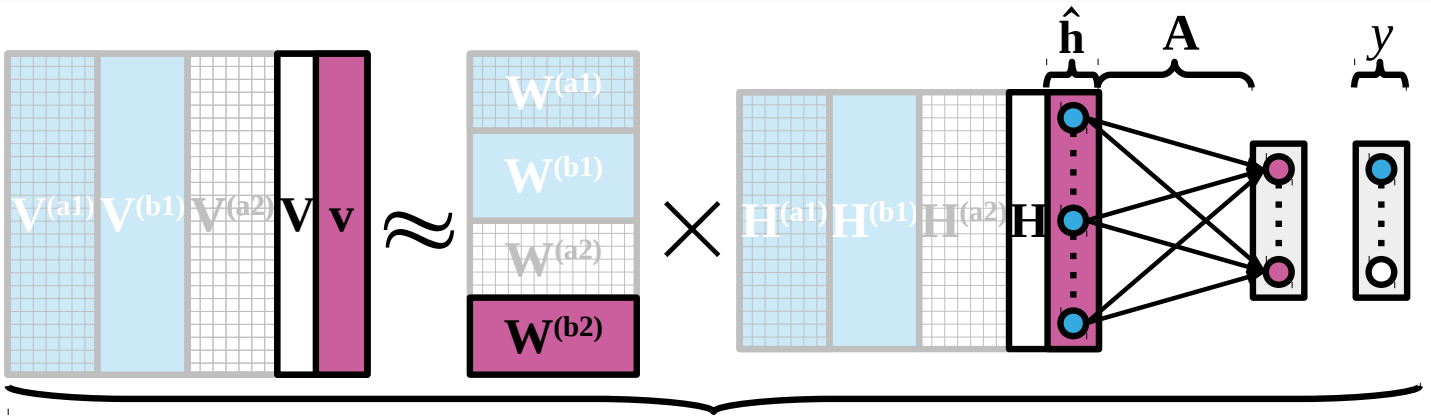
$$\min_{\mathbf{A}} E_{y,\mathbf{v}}[l_s(y,\mathbf{A},\hat{\mathbf{h}}(\mathbf{v},\mathbf{W}))]+\frac{\nu}{2}\|\mathbf{A}\|_2^2$$

Source code is available at `https://github.com/rserizel/TGNMF`

Task-driven group NMF : algorithm (3)

- For each new sample ($\mathbf{v}$)

Update the corresponding dictionnary $\mathbf{W}^{(cs)}$



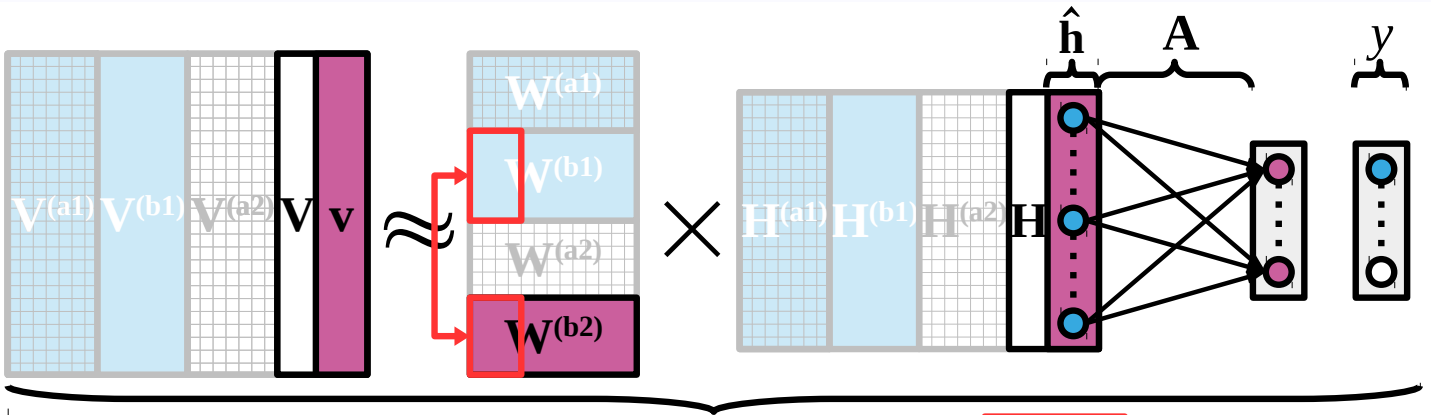
$$\min_{\mathbf{W}^{(cs)}} E_{y,\mathbf{v}}[l_s(y,\mathbf{A},\hat{\mathbf{h}}(\mathbf{v},\mathbf{W}))]+\frac{\nu}{2}\|\mathbf{A}\|_2^2$$

Source code is available at `https://github.com/rserizel/TGNMF`

Task-driven group NMF : algorithm (3)

- **Speaker** similarity constraint

Update the corresponding dictionnary $\mathbf{W}^{(cs)}$



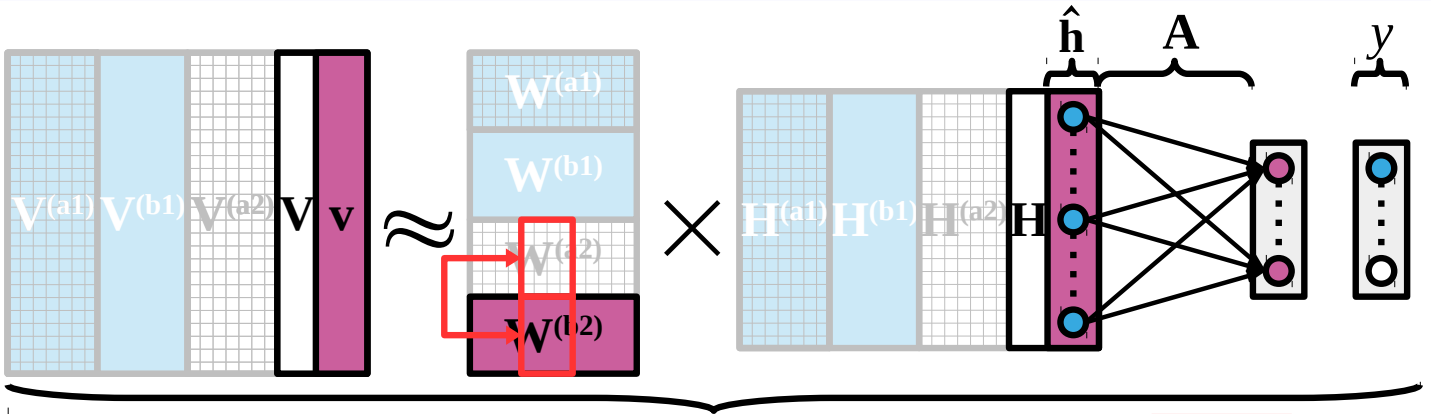
$$\min_{\mathbf{W}^{(cs)}} E_{y,\mathbf{v}}[l_s(y,\mathbf{A},\hat{\mathbf{h}}(\mathbf{v},\mathbf{W}))] + \frac{\nu}{2}\|\mathbf{A}\|_2^2 + \mu_1 J_{\text{SPK}}$$

Source code is available at `https://github.com/rserizel/TGNMF`

Task-driven group NMF : algorithm (3)

- **Session** similarity constraint

Update the corresponding dictionnary $W^{(cs)}$



$$\min_{W^{(cs)}} E_{y,v}[l_s(y, A, \hat{h}(v, W))] + \frac{\nu}{2}\|A\|_2^2 + \mu_1 J_{SPK} + \mu_2 J_{SES}$$

Source code is available at `https://github.com/rserizel/TGNMF`

Experiments

Experiment setup

- Subset of the ESTER corpus ($\approx$ 6 hours training data)
- 132 constant-Q transform coefficients
- Initial dictionary obtained with (group-)NMF : 100 iterations
- Projection on **h** with SPAMS toolbox [a]
- Classifier : multinomial logistic regression
- After 5 epochs : fix **W**, train **A** alone for 50 epochs

a. http ://spams-devel.gforge.inria.fr/

Results (1)

Weighted F1-scores

	Initialisations			
	I-vector	NMF	$GNMF_0$	$GNMF_c$
Unsupervised	76.1%	75.6%	80.7%	81.7%
TNMF Tunning	–	79.9%	81.1%	81.9%

- $GNMF_0$: group NMF **without** similarity constraints
- $GNMF_c$: group NMF **with** similarity constraints (speaker and session)

Results (2)

Weighted F1-scores

		Initialisations	
		$GNMF_0$	$GNMF_c$
Unsupervised		80.7%	81.7%
Tuning	TNMF	81.1%	81.9%
	$TGNMF_0$	81.7%	82.1%
	$TGNMF_c$	82.0%	**82.2%**

- $(T)GNMF_0$: (task-driven) group NMF **without** similarity constraints
- $(T)GNMF_c$: (task-driven) group NMF **with** similarity constraints (speaker and session)

Conclusions and future work

NMF for speaker identification

- Can be competitive with I-vectors

Task-driven NMF

- Large improvements for small dictionaries
- TGNMF$_c$ best performance to date on the corpus

Future work

- Experiment with β-divergence
- Extend the framework to deep learning...

Loria

Further readings

V. Bisot, R. Serizel, S. Essid, and G. Richard. Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification. HAL-archives ouvertes : working paper or preprint (hal-01362864), September 2016. URL `https://hal.archives-ouvertes.fr/hal-01362864`.

V. Bisot, S. Essid, and G. Richard. Overlapping sound event detection with supervised nonnegative matrix factorization. In *Proc. of ICASSP*, 2017.

J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4) :791–804, 2012.

R. Serizel, S. Essid, and G. Richard. Group nonnegative matrix factorisation with speaker and session variability compensation for speaker identification. In *Proc. of ICASSP*, 2016.

Pablo Sprechmann, Alex M Bronstein, and Guillermo Sapiro. Supervised non-euclidean sparse nmf via bilevel optimization with applications to speech enhancement. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, pages 11–15. IEEE, 2014.