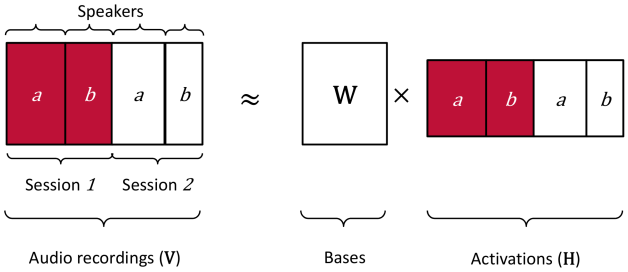
Group NMF with speaker and session variability compensation for speaker identification

Romain Serizel, Slim Essid, Gaël Richard
LTCI, CNRS, Télécom ParisTech, Université Paris - Saclay, 75013, Paris, France

Challenges

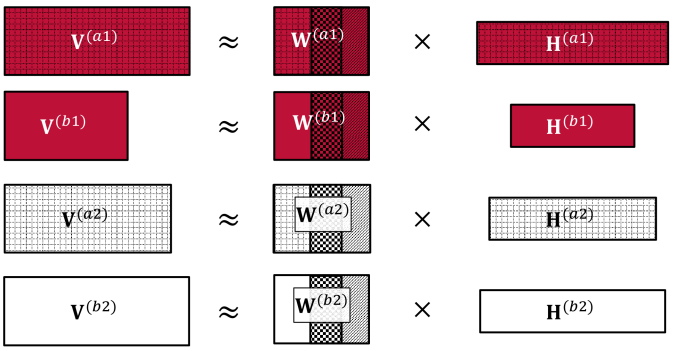
- Feature learning approach based on nonnegative matrix factorisation (NMF) [1]
- Inspired by group NMF [2] and recent work on NMF for speaker identification [3]
- Take speaker and session variability into account as in joint factor analysis and I-vectors [4]

NMF for feature learning in speaker identification



Standard (unsupervised) NMF: $\min_{\mathbf{WH}} D(\mathbf{V}|\mathbf{WH})$ s.t. $\mathbf{W} \geq 0,\ \mathbf{H} \geq 0$ (1)

Class and session similarity constraints



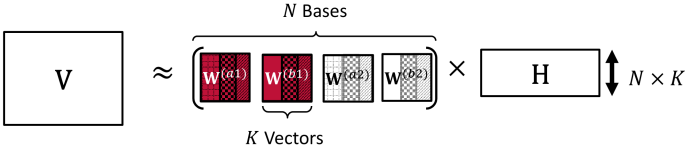
Speaker distance: $J_{\text{SPK}} = \frac{1}{2}\sum_{c=1}^{C}\sum_{s\in\mathcal{S}_c}\sum_{\substack{s_1\in\mathcal{S}_c\\ s_1\neq s}} \|\mathbf{W}_{\text{SPK}}^{(cs)} - \mathbf{W}_{\text{SPK}}^{(cs_1)}\|^2 < \alpha_1$ (2)

Session distance: $J_{\text{SES}} = \frac{1}{2}\sum_{s=1}^{S}\sum_{c\in\mathcal{C}_s}\sum_{\substack{c_1\in\mathcal{C}_s\\ c_1\neq c}} \|\mathbf{W}_{\text{SES}}^{(cs)} - \mathbf{W}_{\text{SES}}^{(c_1 s)}\|^2 < \alpha_2$ (3)

$$\min_{\mathbf{W},\mathbf{H}} \sum_{c=1}^{C}\sum_{s\in\mathcal{S}_c} D_{KL}(\mathbf{V}^{(cs)}|\mathbf{W}^{(cs)}\mathbf{H}^{(cs)}) + \lambda_1 J_{\text{SPK}} + \lambda_2 J_{\text{SES}} \quad \text{s.t.} \quad \mathbf{W} \geq 0,\ \mathbf{H} \geq 0 \tag{4}$$

Feature extraction

Concatenate basis and apply standard NMF.

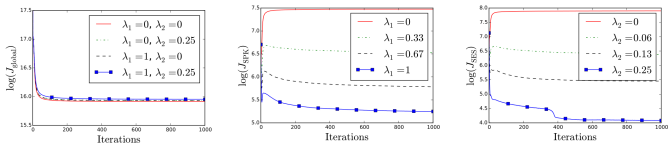


Experiments

- Validated on a subset of the ESTER dataset [5]
  - Train ≈ 6h, Test ≈ 3h,
  - 95 speakers, $N = 236$ unique couples (speaker, session),
  - 132 constant Q transform (CQT) coefficients,
  - $K = 8$ ($K_{\text{SPK}} = 4$, $K_{\text{SES}} = 2$, $K_{\text{RES}} = 2$).
- Speakers repartition according to training data:

Duration	< 1min	1min – 5min	> 5min
Number of speakers	25	26	44

- Convergence of the different criteria.



(a) KL-divergence (b) Intra-speaker distance across sessions($\lambda_2 = 0$) (c) Intra-session distance across speakers ($\lambda_1 = 0$)

- F1-scores with multinomial logistic regression.

Features	I-vector	NMF	Group-NMF $\lambda_1 1 = 0$ $\lambda_2 = 0$	Group-NMF $\lambda_1 = 0.33$ $\lambda_2 = 0.06$
F1-score	76.1%	70.7%	77.8%	**80.2%**

Conclusions

- Competitive with I-Vector
- Improve NMF performance
- More extensive tests to be performed

References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] H. Lee and S. Choi, "Group nonnegative matrix factorization for EEG classification," in *Proc. of AISTATS*, 2009, pp. 320–327.

[3] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Similarity induced group sparsity for non-negative matrix factorisation," in *Proc. of ICASSP*, 2015, pp. 4425–4429.

[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] G. Gravier, J. F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait and K. Choukri, "ESTER, une campagne d'évaluation des systemes d'indexation automatique d'émissions radiophoniques en français.," in *Proc. of Journées d'Etude sur la Parole*, 20041.

This work was partly funded by the European Union under the FP7-LASIE project (grant 607480).

romain.serizel@telecom-paristech.fr

https://github.com/rserizel/groupNMF

ICASSP - March 2016